

# Social Networks of the Past: Information Extraction from Historical Demographic Documents

Josep Lladós



Wurzburg, February 2016

# Index

- Motivation and context
- Word spotting
  - Query by example
  - Query by string
- Context aware word spotting
- Record linkage
- Conclusions



# Index

- **Motivation and Context**
- Word Spotting
  - Query by example
  - Query by string
- Context Aware Word spotting
- Record linkage
- Conclusions

# The impact of historical genealogical data

- Although digitization campaigns, structured knowledge from document archives is still inaccessible by computers and humans.
- These documents reflect the identity of the past. Unlocking their contents allows citizens to know the collective and evolving memory of their society.
- Demographic (Birth, Dead, Marriage, Census) are snapshots of societies and their socio-economic evolution.
- Massive extraction of such data would result in a huge “social network” of the past, so data analytics techniques currently used in social media research would be a powerful tool for historians and social scientists.
- But ...

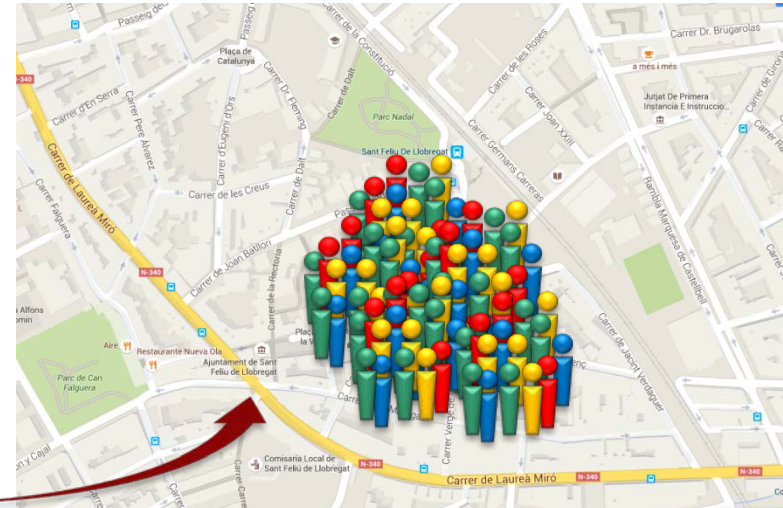
**... fast and efficient data entry and interpretation tools are required.**



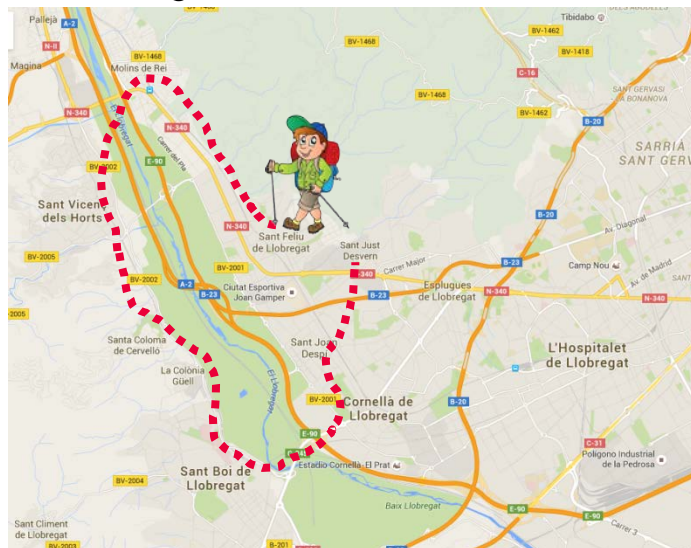
# Innovative services can be envisioned



Time  
machine

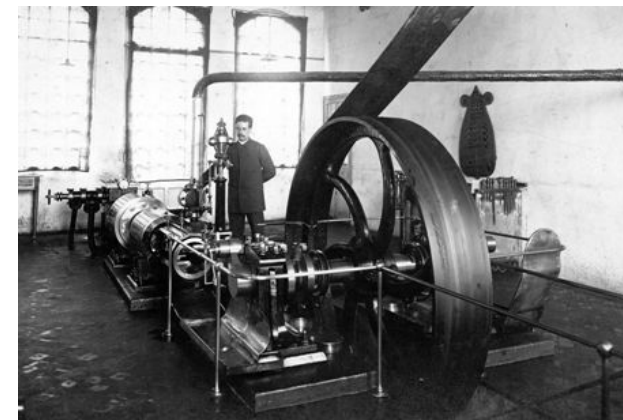


## Genealogic tourism



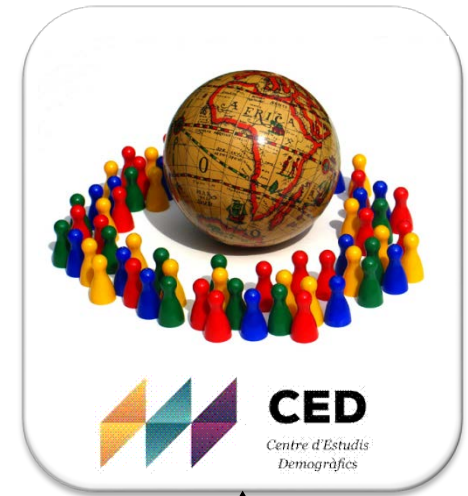
Genealogies

## Impact of industrialization





# Ongoing projects CED and CVC



- Multidisciplinary team:
  - **Engineers:** Centre de Visió per Computador
  - **Social Scientists:** Centre d'Estudis Demogràfics
- To develop tools and procedures to facilitate the massive digitalization of demographic sources from the past to the construction of public databases and the improvement and search of the archives' documents.
- End-to-end process: from the recognition of the images to the interpretation of the databases.

# The Project Five Centuries of Marriages



- Advanced Grant – European Research Council.
- 2011 – 2016.
- Partners:
  - Universitat Autònoma de Barcelona (UAB).
  - Centre d'Estudis Demogràfics (CED).
  - Centre de Visió per Computador (CVC).
- Aim:



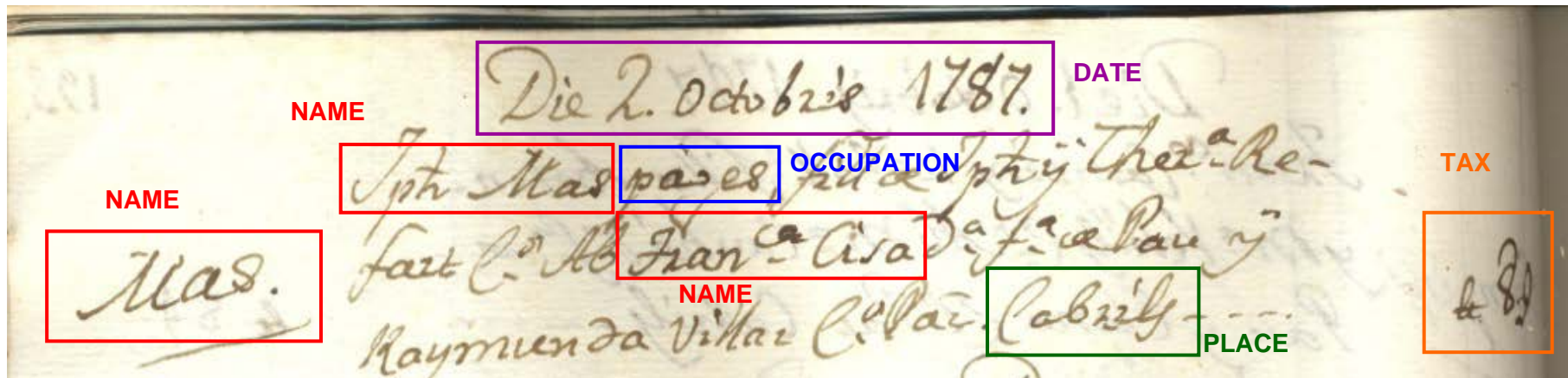
This project is a long-term research initiative based on the data-mining of the *Llibres d'Esposalles* conserved at the Archive of the Barcelona Cathedral. This extraordinary data source comprises 291 books of marriage licenses records, with information of approximately 610.000 unions celebrated in over 250 parishes of the Diocese between 1451 and 1905.

# The Barcelona Marriage Licenses



## Old Marriage Licenses of the Cathedral of Barcelona (Spain)

- Pope Benedict XIII established in 1408 a marriage fee (for building the Cathedral)
- 244 books (15th -19th centuries: 1451 to 1905)
- Approx. 614.000 marriage licenses from 250 parish churches (1900).
- The books include information on the couples, their parents, their occupations, and the tax paid depending on their social class.





Source: Barcelona Historical Marriage Licenses

# Index

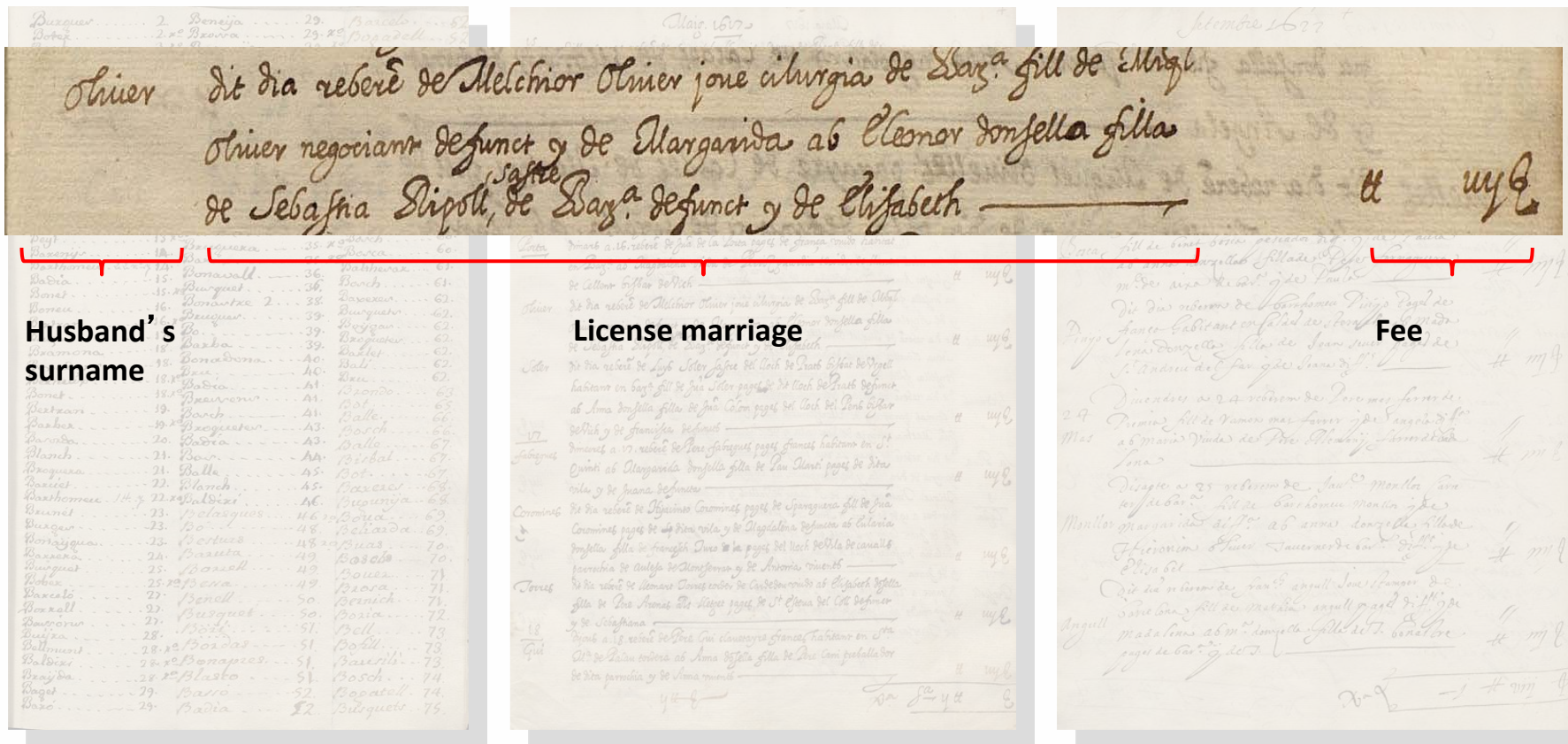
# Marriage Licenses

Buquer	2.	Benija	29.	Bardo	57.
Borax	2. <sup>re</sup>	Berona	29. <sup>re</sup>	Bogardell	58.
Borax	3. <sup>re</sup>	Burguya	29. <sup>re</sup>	Baso	58.
Borax	5.	Balmaceda	31.	Bolvena	54.
Borax	5. <sup>re</sup>	Borgada	31.	Bo Barro	54.
Borax	7.	Borax	31.	Bo Buada	54.
Borax	7.	Borax	32.	Bell	54.
Borax	7. <sup>re</sup>	Borax	32.	Bell	54.
Borax	8.	Borax	32.	Bell	54.
Borax	8. <sup>re</sup>	Borax	32.	Bell	54.
Borax	8. <sup>re</sup>	Borax	32.	Bell	54.
Borax	9.	Borax	32.	Bell	54.
Borax	9. <sup>re</sup>	Borax	32.	Bell	54.
Borax	10.	Borax	32.	Bell	54.
Borax	10. <sup>re</sup>	Borax	32.	Bell	54.
Borax	11.	Borax	32.	Bell	54.
Borax	11. <sup>re</sup>	Borax	32.	Bell	54.
Borax	12.	Borax	32.	Bell	54.
Borax	12. <sup>re</sup>	Borax	32.	Bell	54.
Borax	13.	Borax	32.	Bell	54.
Borax	13. <sup>re</sup>	Borax	32.	Bell	54.
Borax	14.	Borax	32.	Bell	54.
Borax	14. <sup>re</sup>	Borax	32.	Bell	54.
Borax	15.	Borax	32.	Bell	54.
Borax	15. <sup>re</sup>	Borax	32.	Bell	54.
Borax	16.	Borax	32.	Bell	54.
Borax	16. <sup>re</sup>	Borax	32.	Bell	54.
Borax	17.	Borax	32.	Bell	54.
Borax	17. <sup>re</sup>	Borax	32.	Bell	54.
Borax	18.	Borax	32.	Bell	54.
Borax	18. <sup>re</sup>	Borax	32.	Bell	54.
Borax	19.	Borax	32.	Bell	54.
Borax	19. <sup>re</sup>	Borax	32.	Bell	54.
Borax	20.	Borax	32.	Bell	54.
Borax	20. <sup>re</sup>	Borax	32.	Bell	54.
Borax	21.	Borax	32.	Bell	54.
Borax	21. <sup>re</sup>	Borax	32.	Bell	54.
Borax	22.	Borax	32.	Bell	54.
Borax	22. <sup>re</sup>	Borax	32.	Bell	54.
Borax	23.	Borax	32.	Bell	54.
Borax	23. <sup>re</sup>	Borax	32.	Bell	54.
Borax	24.	Borax	32.	Bell	54.
Borax	24. <sup>re</sup>	Borax	32.	Bell	54.
Borax	25.	Borax	32.	Bell	54.
Borax	25. <sup>re</sup>	Borax	32.	Bell	54.
Borax	26.	Borax	32.	Bell	54.
Borax	26. <sup>re</sup>	Borax	32.	Bell	54.
Borax	27.	Borax	32.	Bell	54.
Borax	27. <sup>re</sup>	Borax	32.	Bell	54.
Borax	28.	Borax	32.	Bell	54.
Borax	28. <sup>re</sup>	Borax	32.	Bell	54.
Borax	29.	Borax	32.	Bell	54.
Borax	29. <sup>re</sup>	Borax	32.	Bell	54.

Març. 1672			
Lampis	16	Donat a 15. rebord de francich Lampis fover de Daga fill de francich Lampis fover de Daga y de Maryanna defunta a 6	#
Vingre		Eulonia donyella filla de Juan Soto pagador de Daga y de Arago	#
		De dia rebord de Roman Vingre fover francich vinda hastant ora	
		1 <sup>a</sup> Vicent dels Ports de Paula vinda de Josepha Sota rebord	#
		morí en la Doca	
Samon		De dia rebord de Jaime Ramon pagat hastant en 1 <sup>a</sup> Daga fill de Juan Ramon pagat francich hastant en 1 <sup>a</sup> Daga y de Hieronyma ab	
		Eulonia donyella filla de Juan Plantes pagat negu de cage de	#
		dia vinda y de Juana	#
Porta	16	Donat a 16. rebord de Julia de la Porta pagat de fronsa vinda hastant en Daga ab Magdalena vinda de Perot guardia hoodor de Llana	#
		en Daga filla de Daga	#
Oliver		De dia rebord de Melchor Oliver jma vinda de Daga fill de Melchor Oliver negociant defunta y de Margarida ab Eleanor donyella filla de Josepha Oliver de Daga defunta y de Elisabet	#
Soler		De dia rebord de Luis Soler fover del loch de Trast de Trast de Vagel hastant en Daga fill de Juan Soler pagador de loch de Trast defunta ab Ama donyella filla de Juan Olon pagat del loch del Pont d'Alar	#
		defunta y de francich defunta	#
fatigues	17	Donat a 17. rebord de Pere fatigues pagat francich hastant en 1 <sup>a</sup> Quinti ab Margarida donyella filla de Luis Oloni pagat de dita vila y de Juana defunta	#
Corominas		De dia rebord de Jhsuinas Corominas pagat de Sparaguera fill de Juan Corominas pagat de dita vila y de Magdalena defunta ab Eulonia donyella filla de francich Daga y de pagat del loch de la casa de cavallis	#
		hastant de Alegha de Almagro y de Antonia vinda	#
Torres		De dia rebord de Ramon Torres enter de Capdesordenat ab Elisabet donyella filla de Jose Frances de Arago pagat de 1 <sup>a</sup> Elena del Cell fover	#
		de Josepha	#
Gui	18	Donat a 18. rebord de Pere Gui negociant francich hastant en 1 <sup>a</sup> Daga y de Paula vinda ab Ama donyella filla de Jose Ceni prebador de dita parroquia y de Anna vinda	#
		ya f	
		ya f	

[illegible]

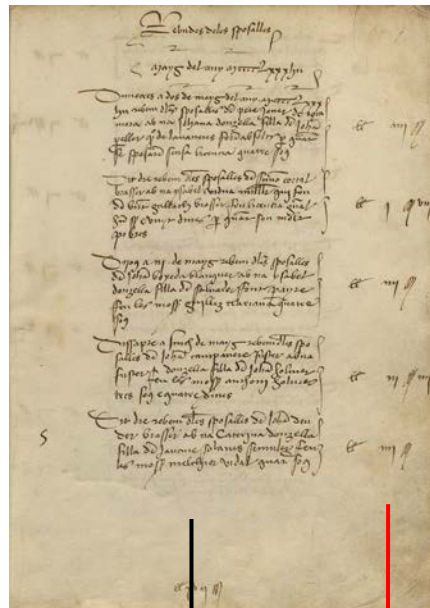
# Source: Barcelona Historical Marriage Licenses





# Source: Continuity

1481: volume 3

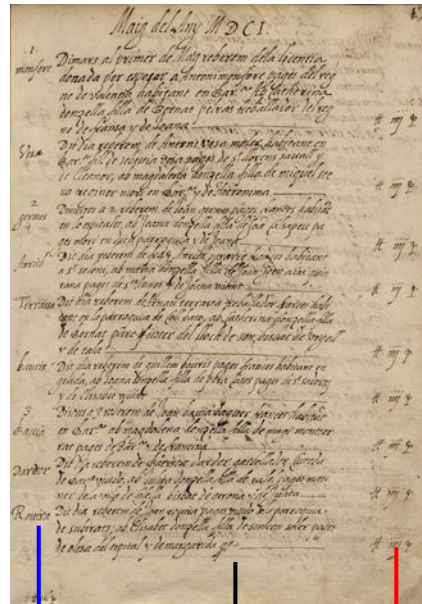


Marriage license

Husband's surname

Fee

1601: volume 61

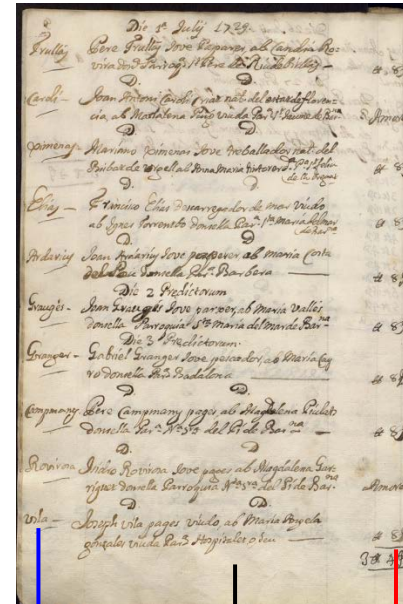


Marriage license

Husband's surname

Fee

1729: volume 127

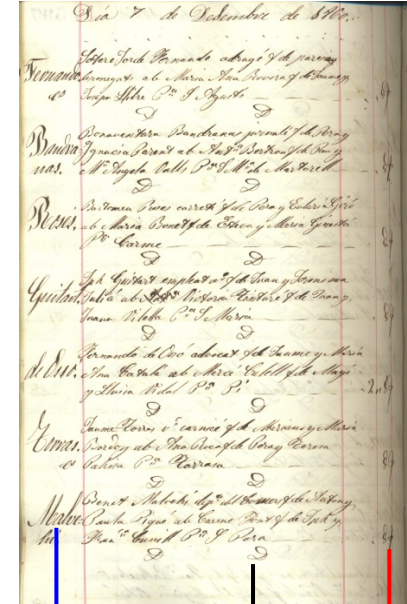


Marriage license

Husband's surname

Fee

1860: volume 200



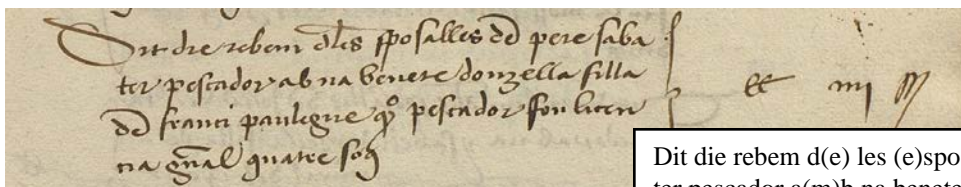
Marriage license

Husband's surname

Fee

# Source: Contents

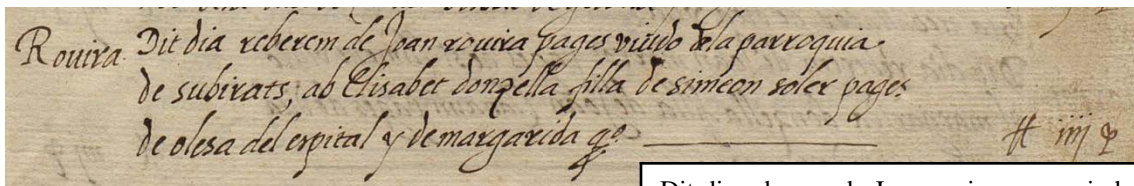
## 1481: volume 3



Dit die rebem d(e) les (e)sposalles de pere saba /  
ter pescador a(m)b na benete donzella filla /  
de franci paulegue q(uondam) pescador fou licen /  
tia g(e)n(er)al quatre sous IIII sous

On said date we receive the marriage license fee  
for pere sabater fisherman with maiden benete  
daughter of franci paulegue deceased fisherman it  
was a general license charge of 4 sous IIII sous

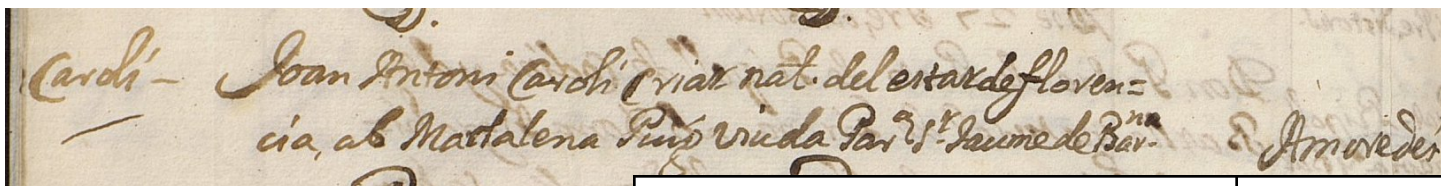
## 1601: volume 61



Dit dia reberem de Joan rouira pages viuud de la  
parroquia / de subirats, a(m)b Elisabet donzella  
filla de simeon soler pages / de olesa del espital y  
de margarida q(uondam) 4 sous

On said date we received from Joan rouira peasant  
widower from the subirats parish and maiden  
Elisabet daughter of simeon soler peasant from olesa  
de espital and the deceased margarida 4 sous

## 1729: volume 127



Joan Antoni Caroli criat nat(ural) del estat de floren-  
cia, a(m)b Marlalena Puig viuuda Par(roquia) s(an)t  
Jaume de Bar(ce)lona Amore dei

Joan Antoni Caroli valet from the state of florence  
with Marlalena Puig widow from sant Jaume of  
Barcelona Parish. Amore dei (no fee)

# The EINES Project (RecerCaixa Program)

- Census records of the municipality of Sant Feliu de Llobregat between 1828 and 1955.
- 2015 – 2017.
- Partners:
  - Centre d'Estudis Demogràfics (CED).
  - Centre de Visió per Computador (CVC).
- Aim:



The EINES project aims to develop **tools and procedures to facilitate the massive digitalization of demographic sources from the past** to the construction of **public databases** and the improvement and search of the archives' documents. We also propose to build some **analytics tools** to generate genealogies with these data, to establish individual and family lifespans and to spatially locate family networks, among other tools.



# Census records of Sant Feliu de Llobregat

NÚMERO		Calle, plaza, poses, caserio, cortijada, etc.	NOMBRES Y APELLIDOS	Fecha del nacimiento	Parentesco o razón de convivencia con el cabeza de familia	NATURALEZA		Profesión, oficio u ocupación	Sexo, estado civil y edad	RESIDENCIA LEGAL		Clasificación vecinal del habitante
De orden	De las hojas					Municipio	Provincia			Municipio	Provincia	
<u>Continúa de Sección 1ª</u>												
1	2	id	Balmacià, Feliu, Pau	18-5-94	hijo	Bombala, Feliu, Pau	Barcelona	maestro	27	Bombala, Feliu, Pau	Barcelona	hijo
2	3	id	Marcos, Feliu, Pau	18-5-94	hijo	Marcos, Feliu, Pau	Barcelona	maestro	27	Marcos, Feliu, Pau	Barcelona	hijo
3	4	id	Marcos, Feliu, Pau	18-5-94	hijo	Marcos, Feliu, Pau	Barcelona	maestro	27	Marcos, Feliu, Pau	Barcelona	hijo
4	5	id	Marcos, Feliu, Pau	18-5-94	hijo	Marcos, Feliu, Pau	Barcelona	maestro	27	Marcos, Feliu, Pau	Barcelona	hijo
5	6	id	Marcos, Feliu, Pau	18-5-94	hijo	Marcos, Feliu, Pau	Barcelona	maestro	27	Marcos, Feliu, Pau	Barcelona	hijo
6	7	id	Marcos, Feliu, Pau	18-5-94	hijo	Marcos, Feliu, Pau	Barcelona	maestro	27	Marcos, Feliu, Pau	Barcelona	hijo
7	8	id	Marcos, Feliu, Pau	18-5-94	hijo	Marcos, Feliu, Pau	Barcelona	maestro	27	Marcos, Feliu, Pau	Barcelona	hijo
8	9	id	Marcos, Feliu, Pau	18-5-94	hijo	Marcos, Feliu, Pau	Barcelona	maestro	27	Marcos, Feliu, Pau	Barcelona	hijo
9	10	id	Marcos, Feliu, Pau	18-5-94	hijo	Marcos, Feliu, Pau	Barcelona	maestro	27	Marcos, Feliu, Pau	Barcelona	hijo
10	11	id	Marcos, Feliu, Pau	18-5-94	hijo	Marcos, Feliu, Pau	Barcelona	maestro	27	Marcos, Feliu, Pau	Barcelona	hijo
11	12	id	Marcos, Feliu, Pau	18-5-94	hijo	Marcos, Feliu, Pau	Barcelona	maestro	27	Marcos, Feliu, Pau	Barcelona	hijo
12	13	id	Marcos, Feliu, Pau	18-5-94	hijo	Marcos, Feliu, Pau	Barcelona	maestro	27	Marcos, Feliu, Pau	Barcelona	hijo

NÚMERO		Calle, plaza, paseo, caserio, cortijada, etc.	NOMBRES Y APELLIDOS	Fecha del nacimiento	Parentesco o razón de convivencia con el cabeza de familia	NATURALEZA		Profesión, oficio u ocupación	Sexo, estado civil y edad	RESIDENCIA LEGAL		Clasificación vecinal del habitante
De orden	De las hojas					Municipio	Provincia			Municipio	Provincia	
1	2	id	Balmacià, Feliu, Pau	18-5-94	hijo	Bombala, Feliu, Pau	Barcelona	maestro	27	Bombala, Feliu, Pau	Barcelona	hijo
2	3	id	Marcos, Feliu, Pau	18-5-94	hijo	Marcos, Feliu, Pau	Barcelona	maestro	27	Marcos, Feliu, Pau	Barcelona	hijo
3	4	id	Marcos, Feliu, Pau	18-5-94	hijo	Marcos, Feliu, Pau	Barcelona	maestro	27	Marcos, Feliu, Pau	Barcelona	hijo
4	5	id	Marcos, Feliu, Pau	18-5-94	hijo	Marcos, Feliu, Pau	Barcelona	maestro	27	Marcos, Feliu, Pau	Barcelona	hijo
5	6	id	Marcos, Feliu, Pau	18-5-94	hijo	Marcos, Feliu, Pau	Barcelona	maestro	27	Marcos, Feliu, Pau	Barcelona	hijo
6	7	id	Marcos, Feliu, Pau	18-5-94	hijo	Marcos, Feliu, Pau	Barcelona	maestro	27	Marcos, Feliu, Pau	Barcelona	hijo
7	8	id	Marcos, Feliu, Pau	18-5-94	hijo	Marcos, Feliu, Pau	Barcelona	maestro	27	Marcos, Feliu, Pau	Barcelona	hijo
8	9	id	Marcos, Feliu, Pau	18-5-94	hijo	Marcos, Feliu, Pau	Barcelona	maestro	27	Marcos, Feliu, Pau	Barcelona	hijo
9	10	id	Marcos, Feliu, Pau	18-5-94	hijo	Marcos, Feliu, Pau	Barcelona	maestro	27	Marcos, Feliu, Pau	Barcelona	hijo
10	11	id	Marcos, Feliu, Pau	18-5-94	hijo	Marcos, Feliu, Pau	Barcelona	maestro	27	Marcos, Feliu, Pau	Barcelona	hijo
11	12	id	Marcos, Feliu, Pau	18-5-94	hijo	Marcos, Feliu, Pau	Barcelona	maestro	27	Marcos, Feliu, Pau	Barcelona	hijo
12	13	id	Marcos, Feliu, Pau	18-5-94	hijo	Marcos, Feliu, Pau	Barcelona	maestro	27	Marcos, Feliu, Pau	Barcelona	hijo

NÚMERO			Calle, plaza, poses, caserio, cortijada, etc.	Hm. de la casa o de la vivienda	NOMBRES Y APELLIDOS		Fecha del nacimiento	Sexo (varón o hembra)	Parentesco o razón de convivencia con el cabeza de familia
De orden	De las hojas	De los parentescos de cada hoja							
165	3	id	And. Bancalito	209	Berisa Pirella Puigventos	Luis	18-12-915	hija	hermana
166	1	id		209	Juan Ant. Cos		18-5-1-918	hijo	hijo
167	3	id		209	Berenguer Pineda Ant. Cos		18-5-1-918	hijo	hijo
168	3	id		211	Berenguer Ant. Cos		18-5-1-918	hijo	hijo
169	1	id		211	Berenguer Ant. Cos		18-5-1-918	hijo	hijo
170	3	id		211	Berenguer Ant. Cos		18-5-1-918	hijo	hijo

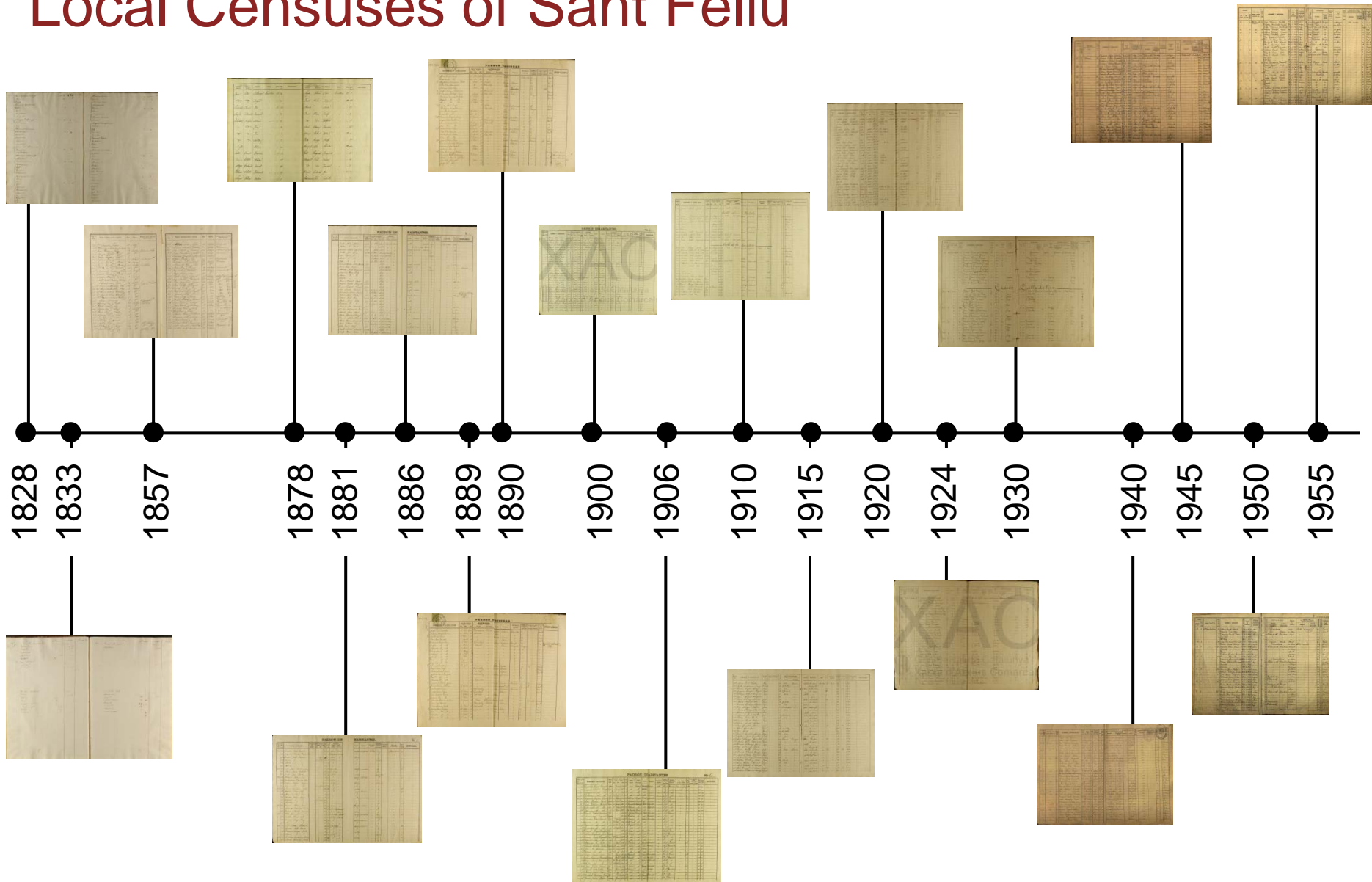
MUNICIPIO DE Sant Feliu de Llobregat PROVINCIA DE Barcelona

**PADRON MUNICIPAL**

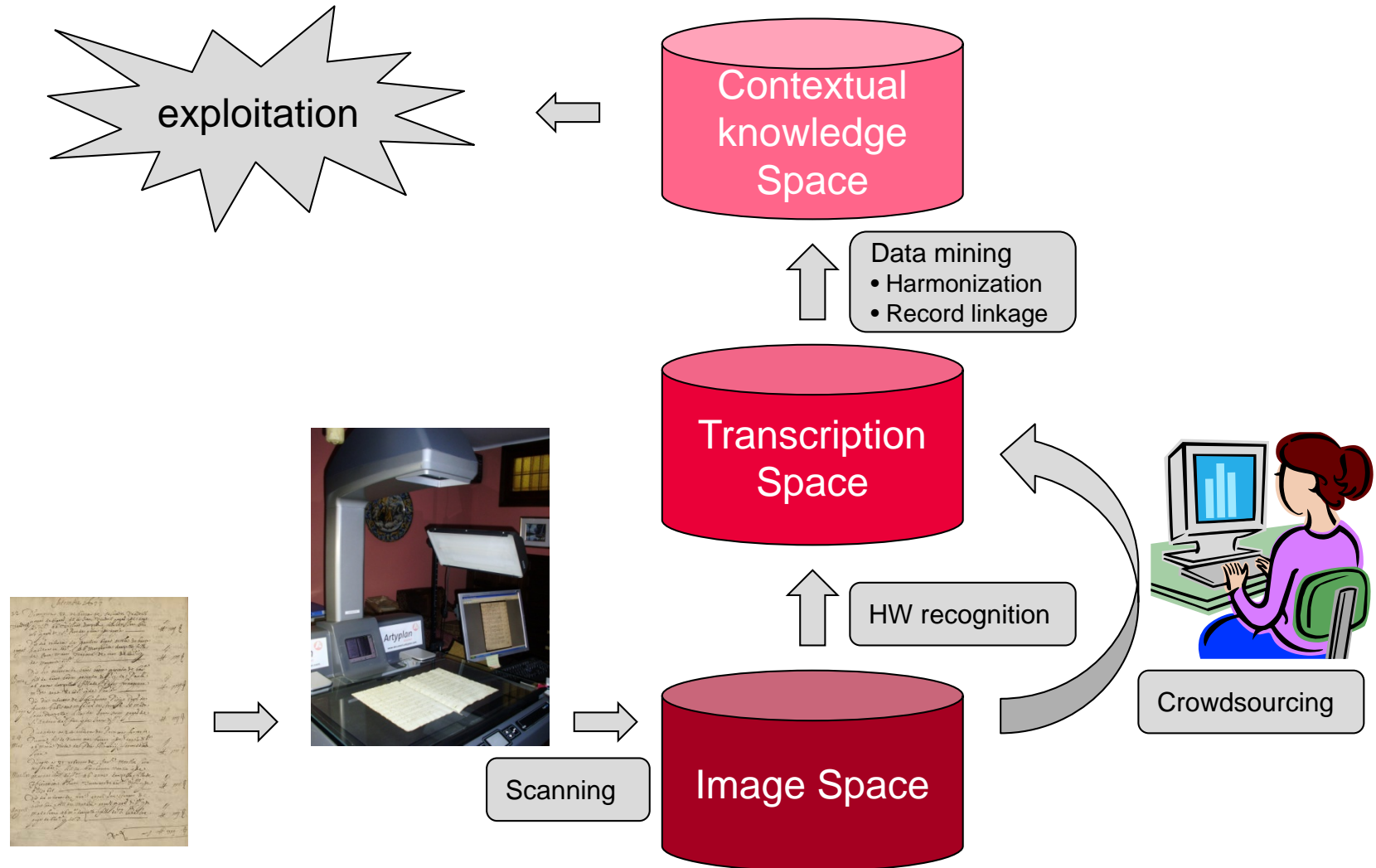
de los vecinos y domiciliados (presentes y ausentes) y transeúntes que se inscribieron en este término el día 31 de Setiembre de 1925

NOTA. — La clasificación vecinal de habitantes de la última casilla, señalada con la nota (a), se hará expresando si el inscrito es cabeza de familia, vecino, domiciliado o transeúnte. En los cabeceras de familia se especificará, además, si son vecinos o transeúntes.

# Local Censuses of Sant Feliu

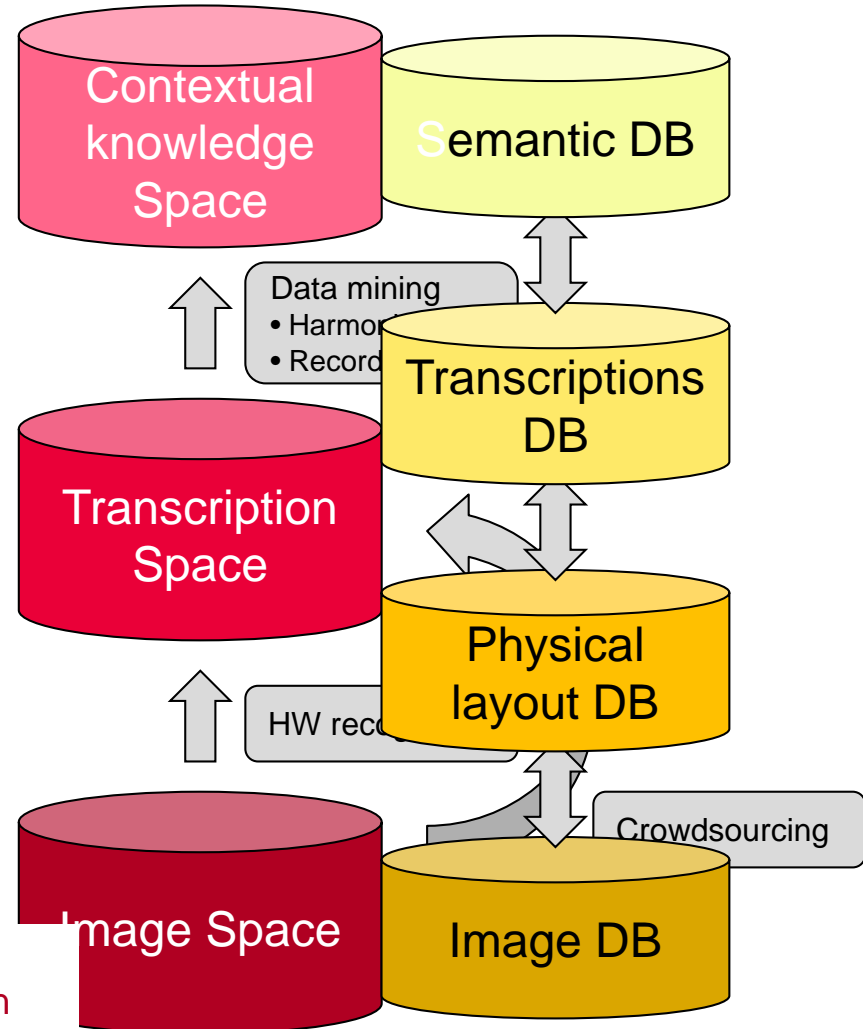
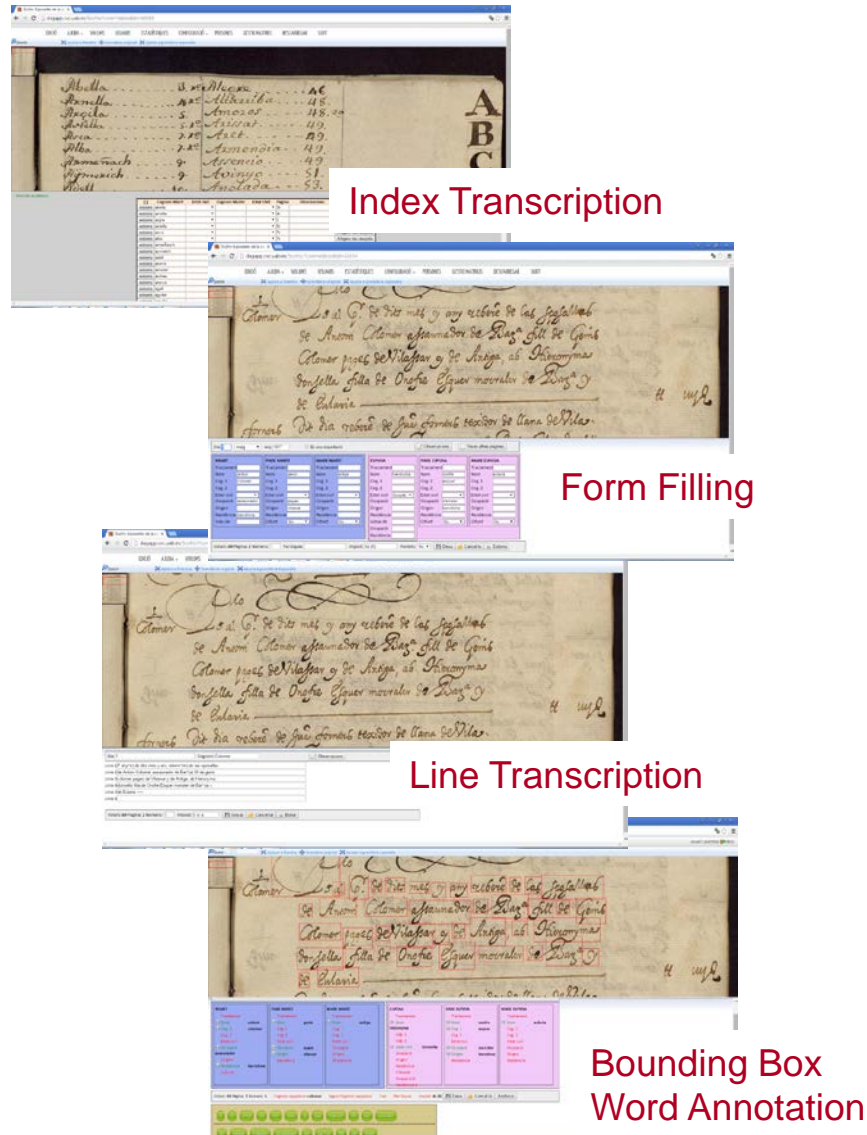


# Technical architecture





# Technical architecture



# Crowdsourcing platform: Form filling

UAB Menú principal Inicial Ajuda Usuaris Volumes Estad. Conf. Persones usuari: aforner desc.

Zoom93% Ajustar a finestra Grandària original Ajustar a grandària esposalla

1 Colomer

al q. de dies mes y any reberé de las esposallas  
de Antoni Colomer assaunador de Barç. fill de Genís  
Colomer pages de Vilassar y de Antiga, ab Hieronyma  
donfella filla de Onofre Esquer morraler de Barç. y  
de Eulària

forner Dit dia reberé de Jua forner texidor de Llana de Vilas.  
Llana habitant en Barç. viudo ab Paula Mas donfella

Dia 1 maig Any 1617 ☐ És una Liquidació Observacions Veure altres pàgines

MARIT	PARE MARIT	MARE MARIT	ESPOSA	PARE ESPOSA	MARE ESPOSA
Tractament	Tractament	Tractament	Tractament	Tractament	Tractament
Nom antoni	Nom genis	Nom antiga	Nom hieronyma	Nom onofre	Nom eularia
Cog. 1 colomer	Cog. 1	Cog. 1	Cog. 1	Cog. 1 esquer	Cog. 1
Cog. 2	Cog. 2	Cog. 2	Cog. 2	Cog. 2	Cog. 2
Estat civil	Estat civil	Estat civil	Estat civil Donzella	Estat civil	Estat civil
Ocupació assaunador	Ocupació pages	Ocupació	Ocupació	Ocupació morraler	Ocupació
Origen	Origen vilassar	Origen	Origen	Origen barcelona	Origen
Residència barcelona	Residència	Residència	Residència	Residència	Residència
Vidu de	Difunt No	Difunt No	Vidua de	Difunt No	Difunt No
			Ocupació		
			Residència		

Volum: 69 Pàgina: 2 Número: 1 Parròquia: Impost: 4s (6) Puntets: No Gravar Cancel·lar Borrar

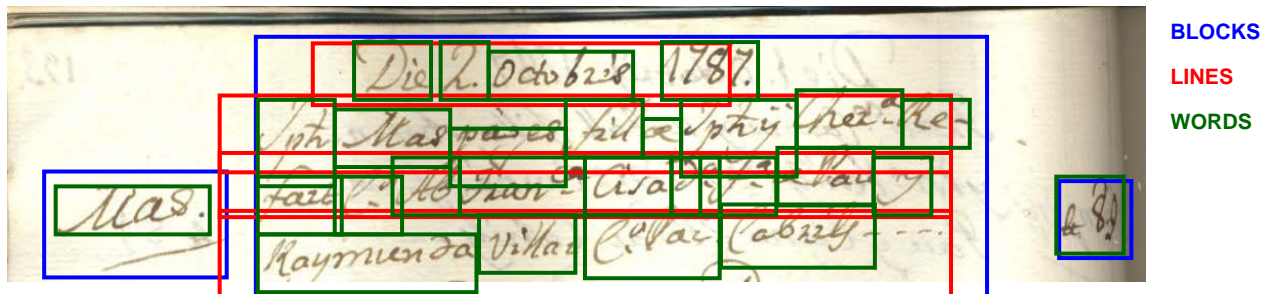


# Index

- Motivation and Context
- Word Spotting
  - Query by example
  - Query by string
- Context Aware Word spotting
- Record linkage
- Conclusions

# Subproblems in the analysis of handwritten images

- **Layout analysis:** to detect (crop) records, lines, words for subsequent recognition.

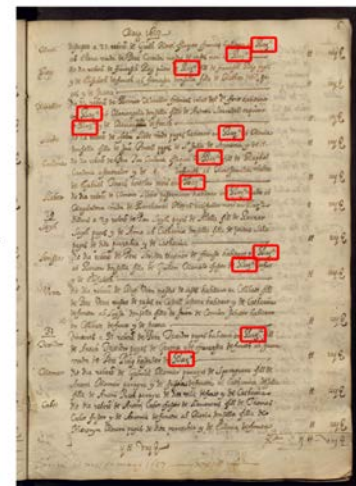
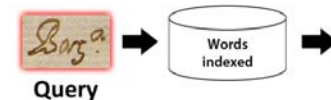


- **Full transcription:** to convert images to editable text.

dit dia reberē de Hieronym Ponsich corder de Bar<sup>a</sup> fill de Jua Pon=

dit dia reberē\$ de Hieronym Ponsich corder de Bar^(a) fill de Jua\$ Pon=

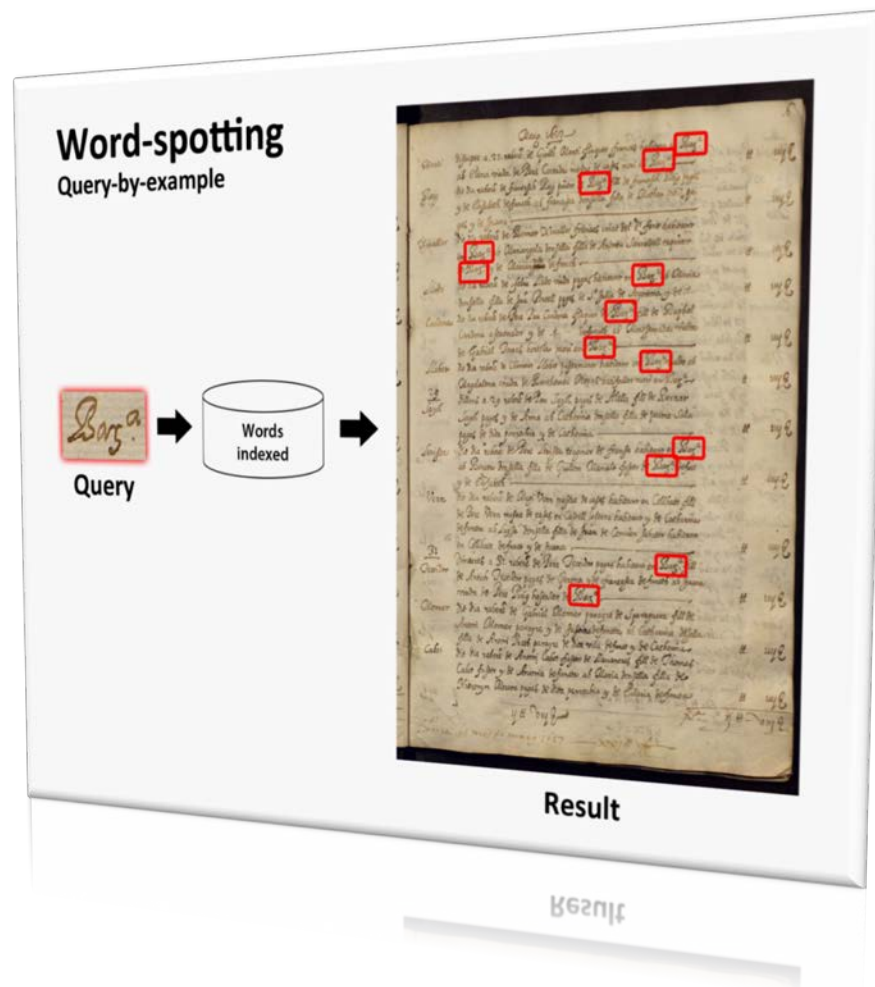
- **Word spotting:** given a query word to search, to locate at image level visually similar word snippets.
  - Query by example
  - Query by string



Result

# Word Spotting: Query-by-Example

- **Objective:** word spotting for indexation and retrieval purposes.
- Given a query word image, we intend to **locate instances** of the same word class into the documents to be indexed, **without explicitly transcribing** the whole document content.
- Words are considered as **shapes**, and spotting is achieved through shape dissimilarity functions.



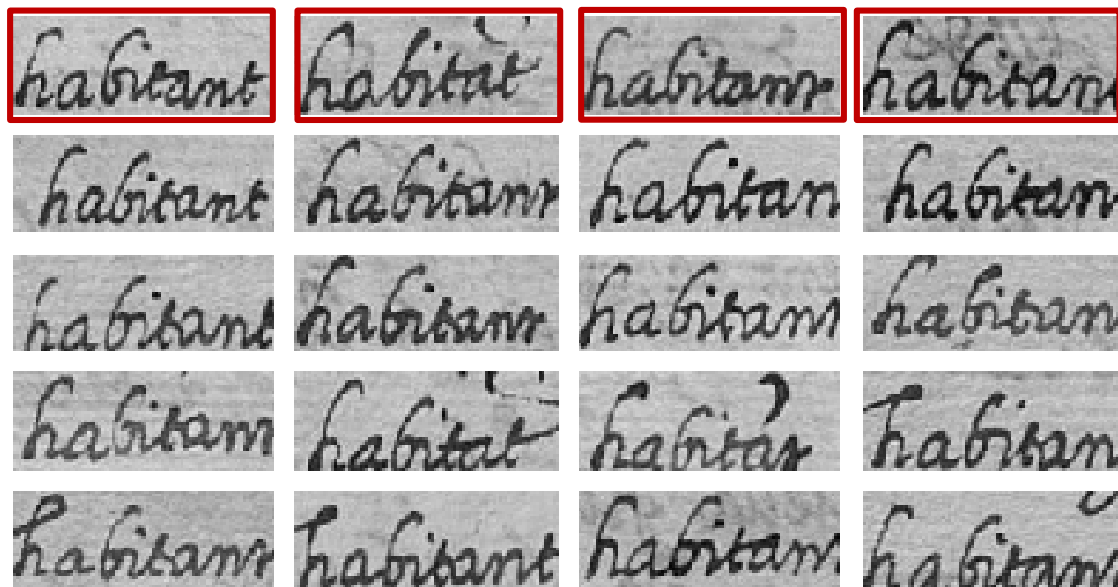
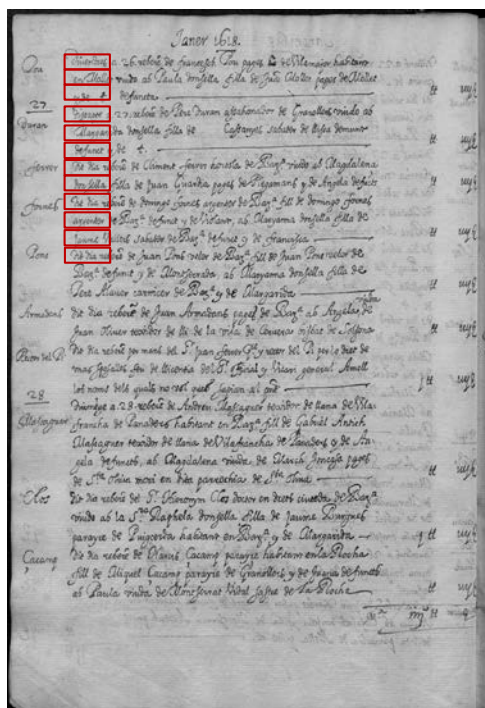
# Cognitive reading

- Word spotting is inspired in the fact that humans tend to read holistically, and recognize the “shape” of the word instead of reading the sentence character-wise.
- Our cognitive system does a remarkable job in making sense out of distorted information when it can put it into context.  
[Graham Rawlinson. *The Significance of Letter Position in Word Recognition*. PhD Thesis, 1976, Nottingham University]

*“Aoccdrnig to a rscheearch at Cmabrigde Uinervtisy, it deosn’t mttar in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rghit pclae. The rset can be a toatl mses and you can sitll raed it wouthit porbelm. Tihs is bcuseae the huamn mnid deos not raed ervey lteter by istlef, but the wrod as a wlohe.”*

# Intuitive idea of query by example word spotting

- Detection using a sliding window over the whole document.





# Word Spotting using a Bag of Visual Words

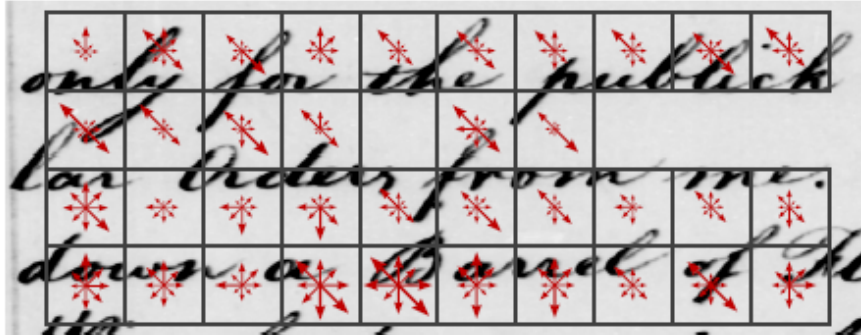
- ◆ Segmentation-free word spotting approach
- ◆ M. Rusiñol, D. Aldavert, R. Toledo and J. Lladós. *Browsing Heterogeneous Document Collections by a Segmentation-free Word Spotting Method*. In Proceedings of the 11<sup>th</sup> International Conference on Document Analysis and Recognition, 2011.

company	company	company,	company	company	company.
English	Company the company	company	company	company	company;
است	hat English sailed an English me	English Ho	in English. “	vo Englishmen	
	y an English	distinguished at	to England I v	in England, an	of England to a
	الله است	ی است	یت است	جد است	من است
	ش است	خیت است	گه است	ی است	زاده است

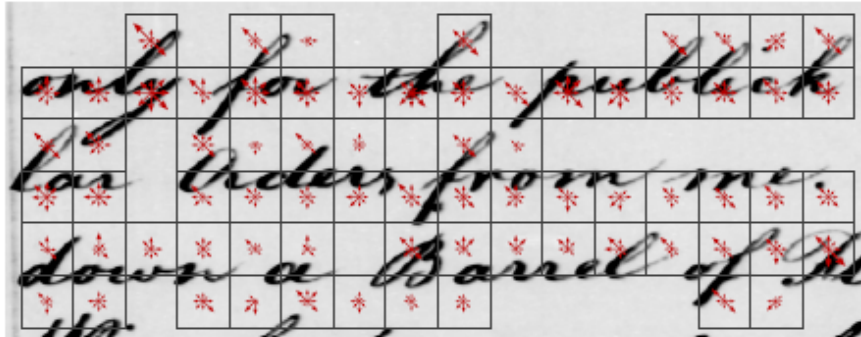
# Word Spotting using a Bag of Visual Words

## *Feature extraction*

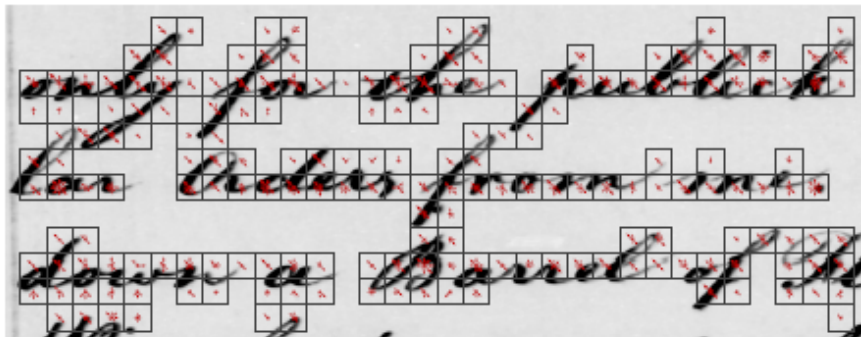
character groups



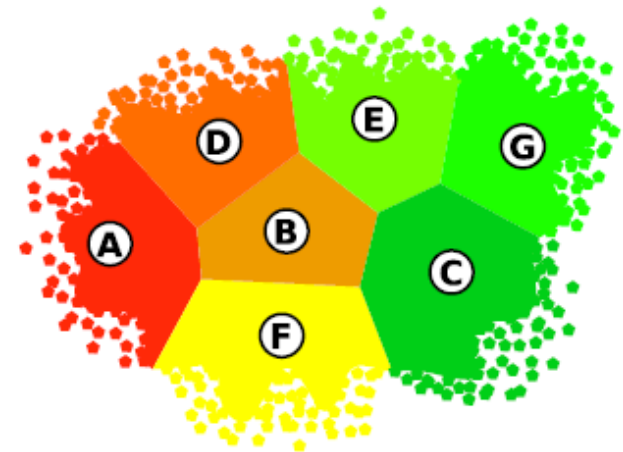
characters



character parts



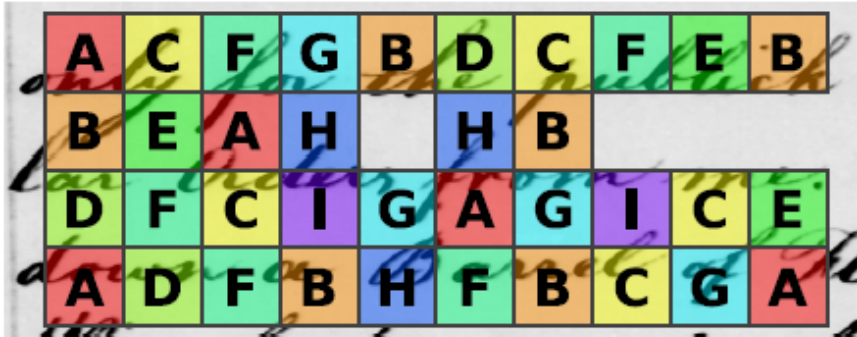
- SIFT features
- Densely extracted at 3 different scales
- Remove low-gradient features
- Build a codebook with *k*-means



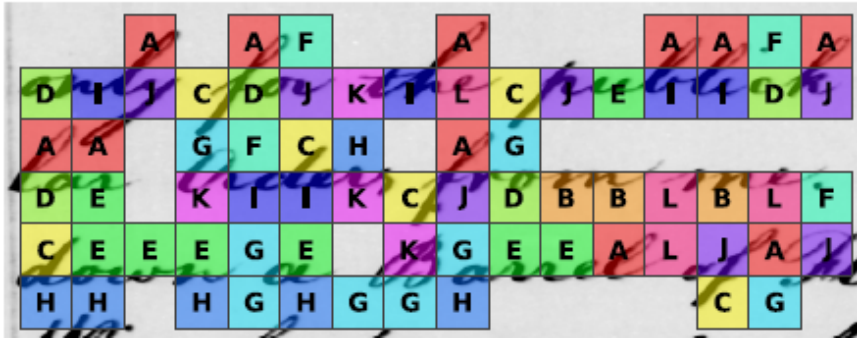
# Word Spotting using a Bag of Visual Words

## *Feature extraction*

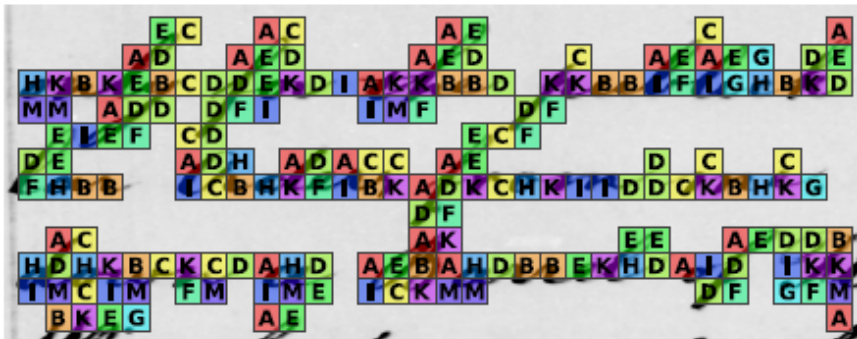
character groups



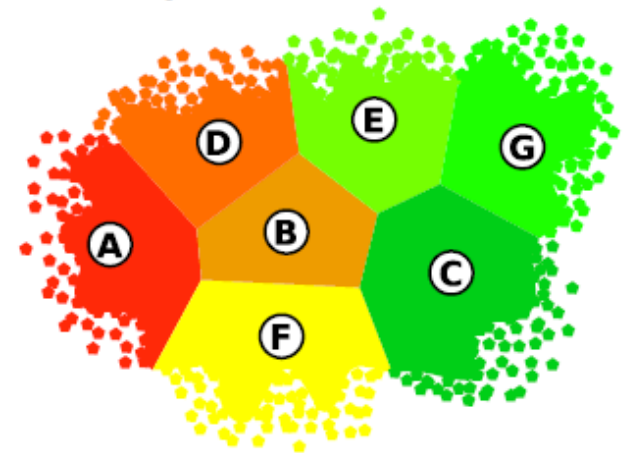
characters



character parts



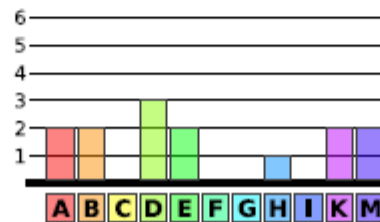
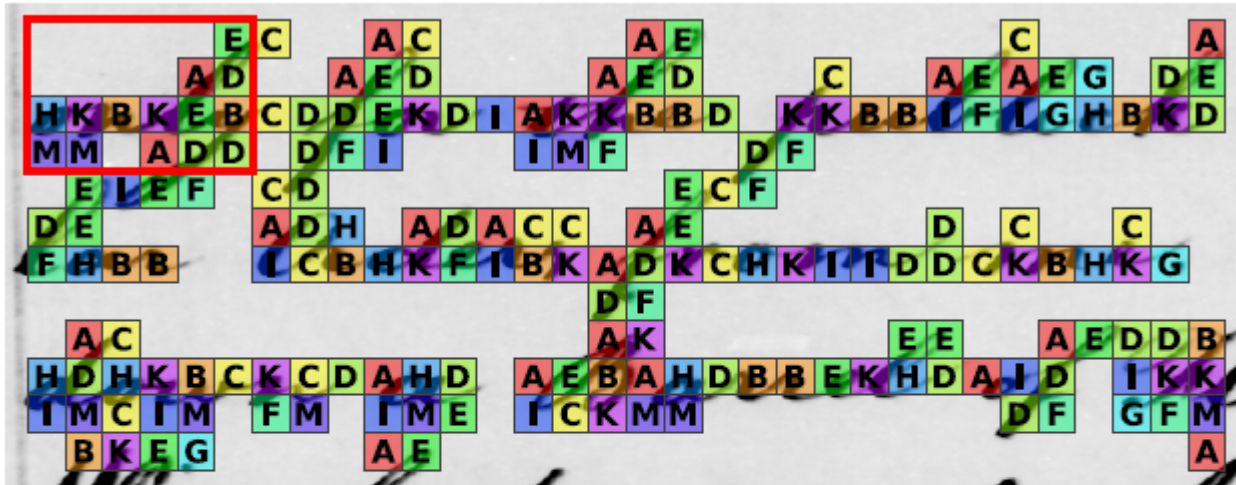
- SIFT features
- Densely extracted at 3 different scales
- Remove low-gradient features
- Build a codebook with *k*-means
- Quantization



# Word Spotting using a Bag of Visual Words

## *Feature extraction*

The document is splitted into overlapping local patches



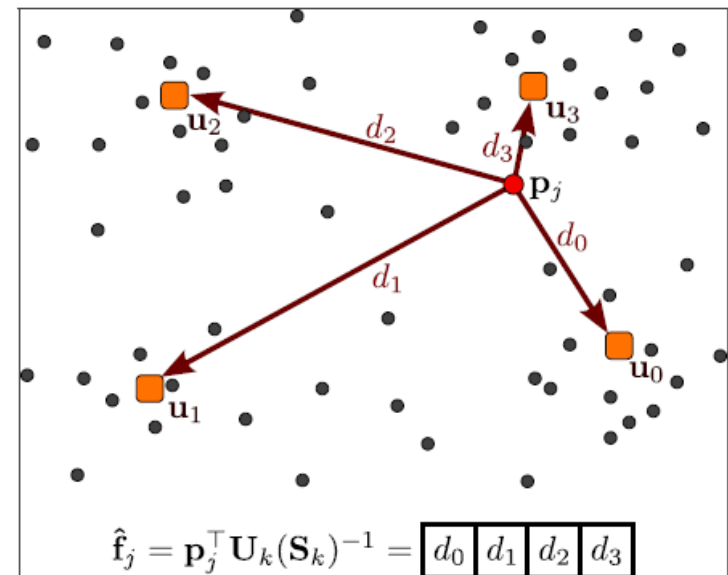
# Word Spotting using a Bag of Visual Words

## Feature extraction

- Final patch descriptor: Latent Semantic Indexing (it projects feature vectors representing patches into a new space of topics with an underlying tf-idf model that emphasizes features that are frequent in a given patch and infrequent in the rest of the corpus).

$$\mathbf{w}_i^T \rightarrow \begin{matrix} & \mathbf{p}_j \\ & \downarrow \\ \begin{bmatrix} f_{1,1} & \dots & f_{1,n} \\ \vdots & \ddots & \vdots \\ f_{m,1} & \dots & f_{m,n} \end{bmatrix} \end{matrix}$$

$$\mathbf{w}_i^T = [f_{i,1} \quad \dots \quad f_{i,n}] \quad \mathbf{p}_j = \begin{bmatrix} f_{1,j} \\ \vdots \\ f_{m,j} \end{bmatrix}$$



$$\mathbf{A} \simeq \hat{\mathbf{A}} = \mathbf{U}_K \mathbf{S}_K (\mathbf{V}_K)^T$$



# Word Spotting using a Bag of Visual Words

## Retrieval

### ◆ Retrieval of patches by similarity



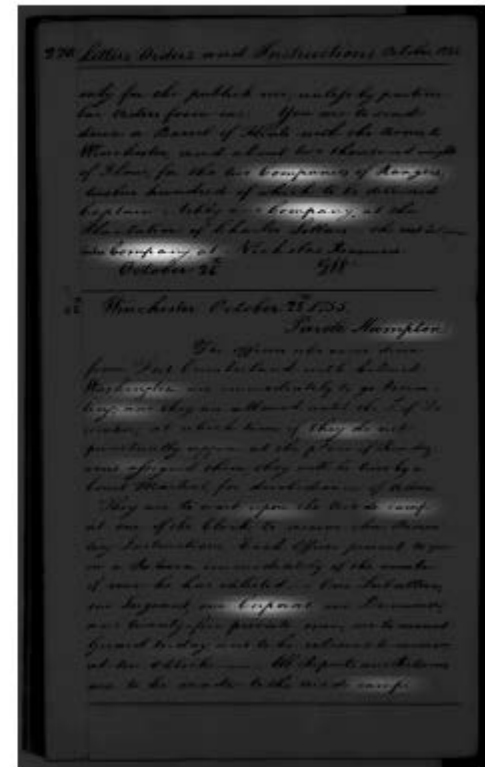
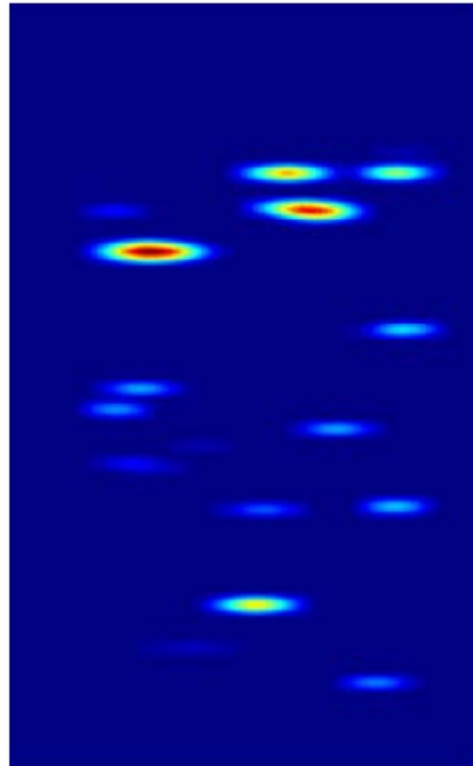
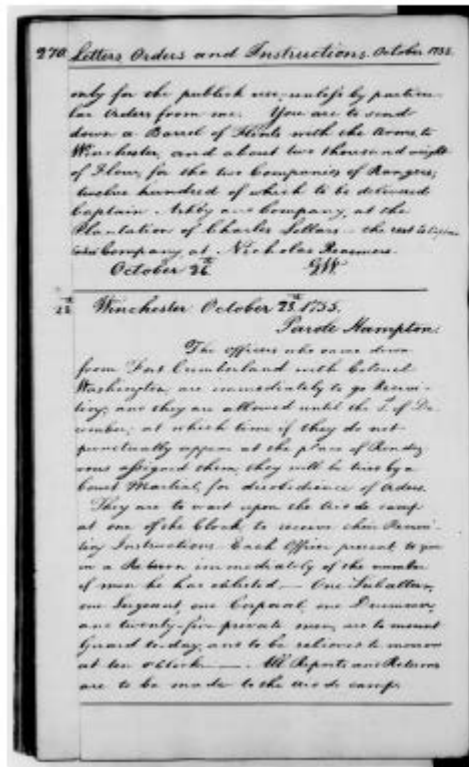


# Word Spotting using a Bag of Visual Words

## Retrieval

- ◆ A voting scheme provides the probability map where to find the queried word

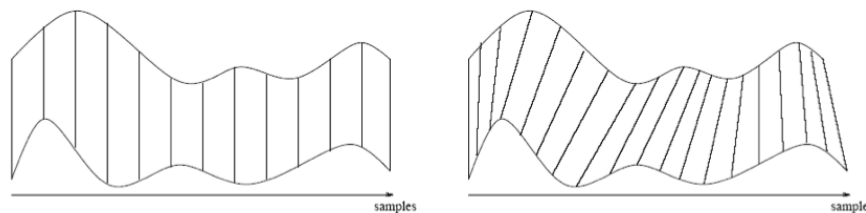
company



# Word Spotting: Results

Baseline → Dynamic Time Warping approach by Rath and Manmatha

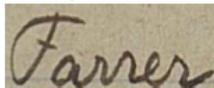
- Features per column
  - Number of foreground pixels
  - Upper profile
  - Lower profile
  - Number of transitions from background to foreground
- Distance
  - Dynamic Time Warping (DTW)



T.M. Rath and R. Manmatha, Word image matching using dynamic time warping, (CVPR'03) 2 (2003), 521-527.

# Word Spotting: Results

Query:



Results:

Farrer	Farrer	Farer	Carrera	Farrer
Farrer	Ferrer	Fuster	Carner	Carassua

Baseline: Dynamic Time Warping

Farrer	Farrer	Farrer	Famade	Farrer
Farrer	Ferrer	Farrer	Farrer	Famand

Bag of Visual Words

Zorra	Faner	Farrer	Ferrer	Riera
Taner	Leiror	Lerna	Ferrer	Riera

Pseudo-structural (LOCI)

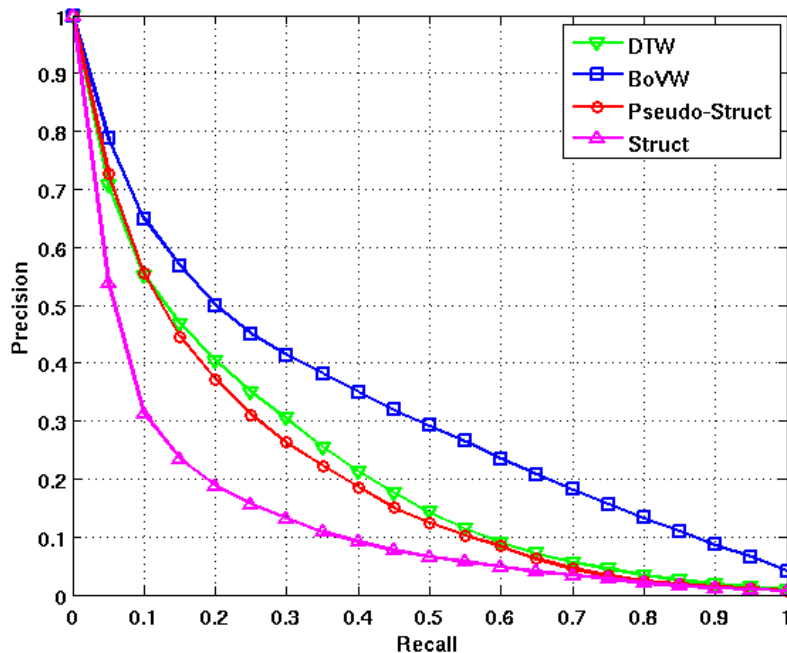
Farrer	Farrer	Ferrer	Farrer	Trassena
Leiror	Taner	Farrer	Zorra	Zora

Structural (Graphs)

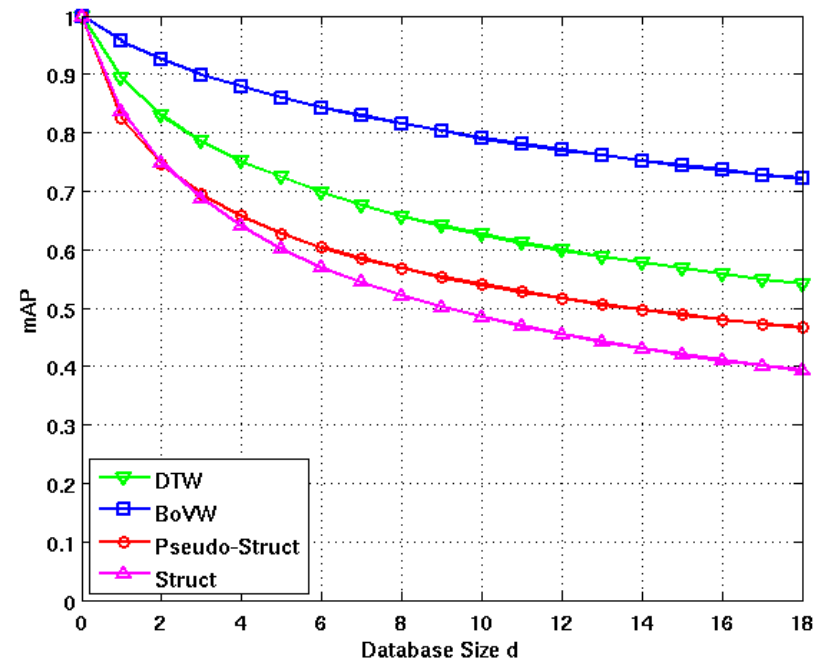
# Word Spotting: Results

- Structural methods are sensitive to noise introduced in the preprocessing step
- DTW is more discriminative but slower
- Bag of Visual Words works better but needs a lot of memory (large vectors)

Precision-Recall



Scalability





# Word Spotting: Results

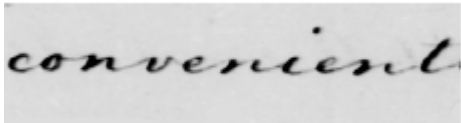
Table 4. Pros and Cons Summary.

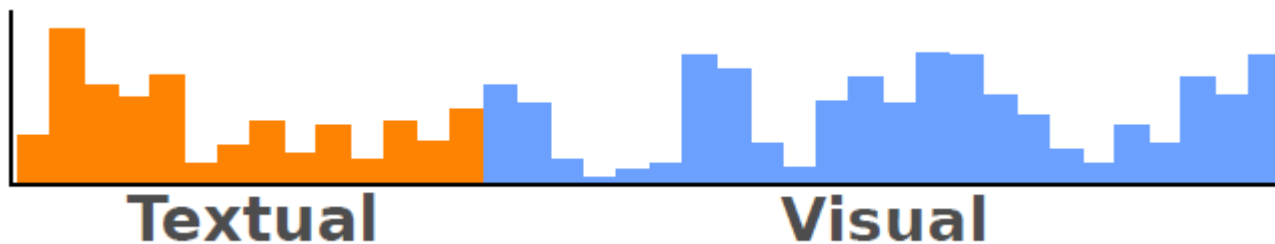
	Size	Time	Indexability	Preprocessing	Performance	Scalability
BoVW	--	+	+	+	+++	— (size)
DTW	+	--	—	—	+	— (time)
Pseudo-structural	+++	+++	+++	—	+	—(discriminative power)
Structural	+++	+++	+++	--	—	— (discriminative power)

# Index

- Motivation and Context
- Word Spotting
  - Query by example
  - **Query by string**
- Context Aware Word spotting
- Record linkage
- Conclusions

# A bi-modal approach

- A word is represented by two information modalities:
  - Text information: convenient
  - Image information: 
- We calculate a **bag-of-features** signature for each information modality.
- The final representation is obtained by **concatenating** both
- signatures into a single **vector**.



# Textual Representation

- Textual information is represented as a combination of uni-grams, bi-grams and tri-grams:

**Text string: convenient**

1-grams

**c** 1 **o** 1 **n** 3 **v** 1 **e** 2 **i** 1 **t** 1

2-grams

**co** 1 **on** 1 **nv** 1 **ve** 1  
**en** 2 **ni** 1 **ie** 1 **nt** 1

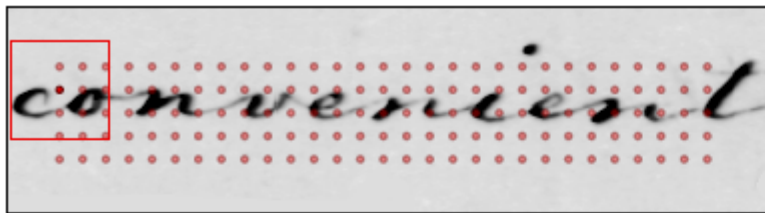
3-grams

**con** 1 **onv** 1 **nve** 1 **ven** 1  
**eni** 1 **nie** 1 **ien** 1 **ent** 1

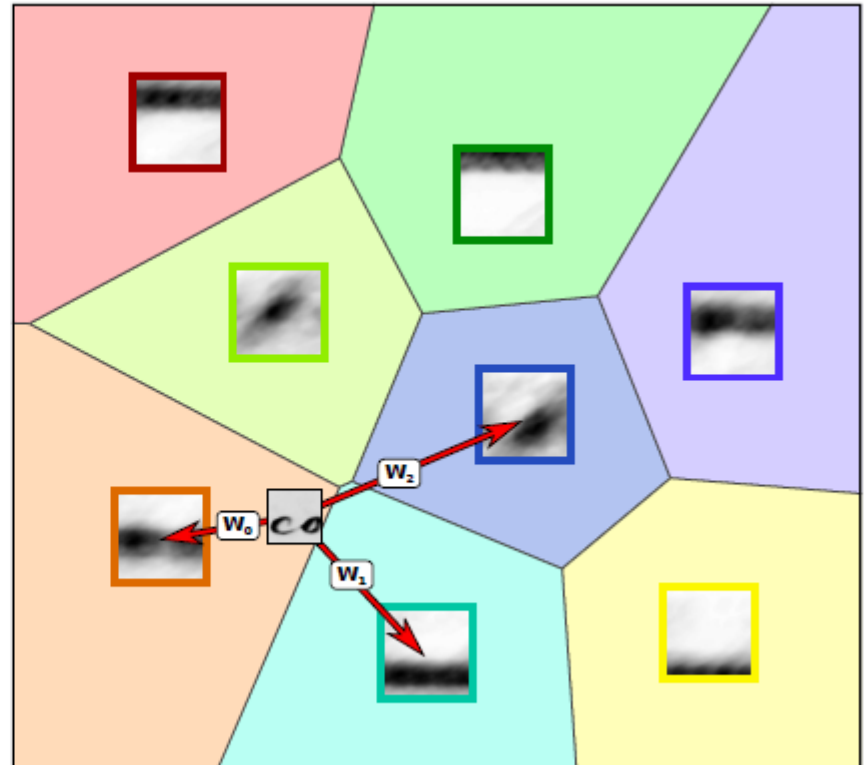
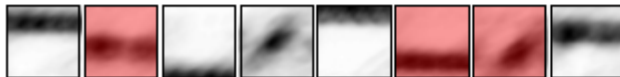


# Visual Representation

- Bag of Visual Words using SIFT descriptor:



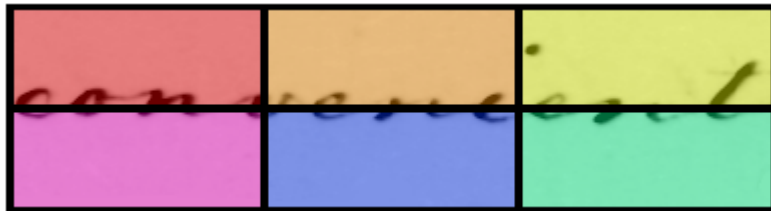
SOFT ASSIGNMENT



# Visual Representation

- We use a two layers spatial pyramid of in order to take into account the spatial distribution of the visual words.

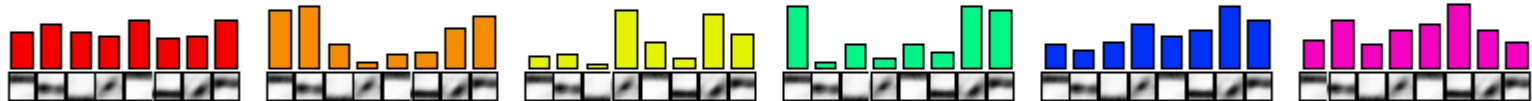
LEVEL 1



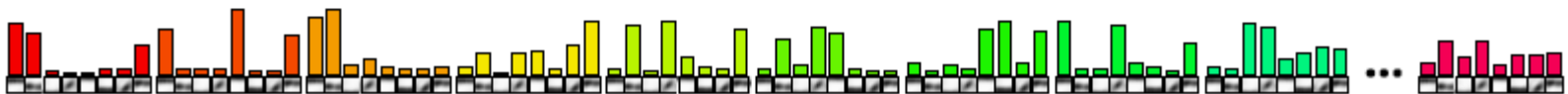
LEVEL 2



LEVEL 1

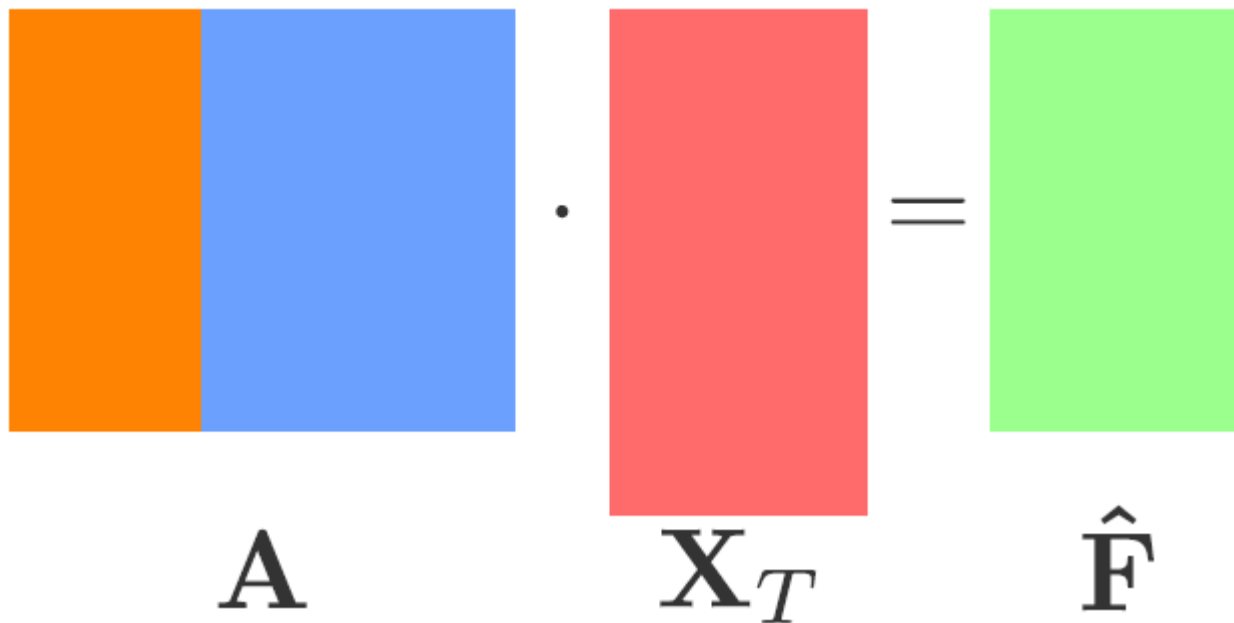


LEVEL 2



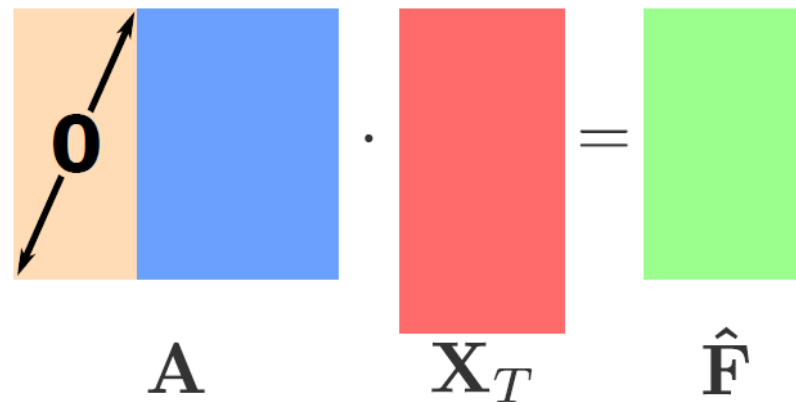
# Multimodal fusion: topics model

- We use Latent Semantic Analysis to infer a topics model from the original features (textual and visual).
- Feature vectors are projected into the topics space by a matrix  $X_T$ .

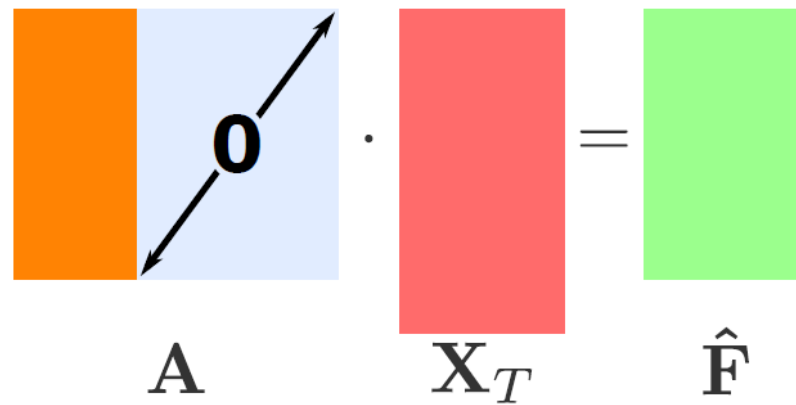

$$A \cdot X_T = \hat{F}$$

# Multimodal fusion: topics model

- The **corpus index** is created using **only visual** information.

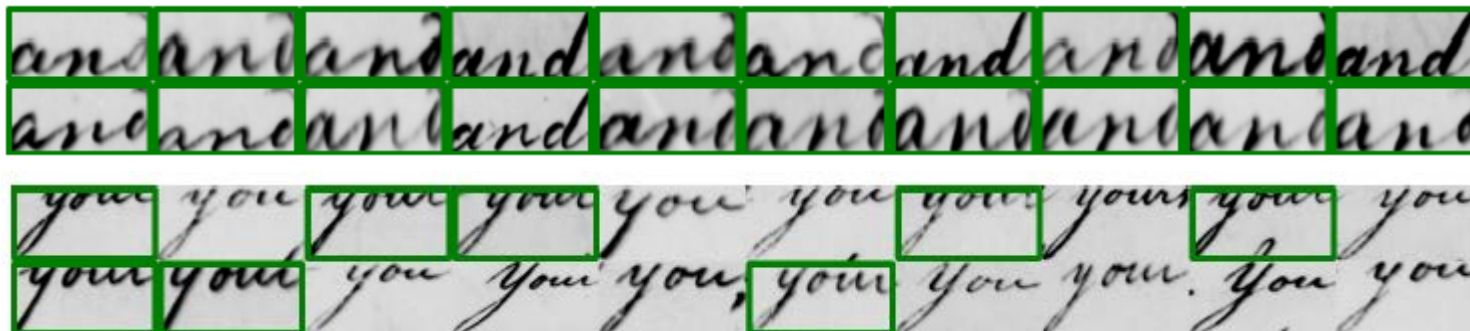


- **Queries** are generated using **only textual** information.





# Results



Method	Segmentation	Cross-validation	Queries	mAP
Proposed	Word-level	4 folds	All words	56.54%
			In-vocabulary words	76.2%
			Liang's 38 queries	83.12%
Liang et al. <sup>5</sup>	Word-level	5 folds	38 queries	67% at rank 10
Fischer et al. <sup>6</sup>	Line-level	4 folds	In-vocabulary words	62.08%
Frinken et al. <sup>7</sup>	Line-level	4 folds	In-vocabulary words	71%

<sup>5</sup>Liang et al. *A synthesised word approach to word retrieval in handwritten documents*. *Patt. Rec.* 45(12):2089–2105, 2009.

<sup>6</sup>Fischer et al. *Lexicon-free handwritten word spotting using character HMMs*. *Patt. Rec.* 33(7):934–942, 2012.

<sup>7</sup>Frinken et al. *A novel word spotting method based on recurrent neural networks* *IEEE TPAMI* 34(2):211–224, 2012.

# Index

- Motivation and Context
- Word Spotting
  - Query by example
  - Query by string
- **Context Aware Word spotting**
- Record linkage
- Conclusions



# The role of the context in visual object recognition

- Classical word spotting is based on the statistics of local terms.
- **Contextual information** can provide more relevant information for the recognition of a query word than intrinsic word image information.
- **Context** of a word in the document can be defined in terms of the other recognized words and their mutual dependences.

# Types of context

- Three types of context can be defined:
  - **word co-occurrence** in a given image segment.
  - **geometric context** involving a language model regarding to the relative 1D or 2D position of objects → visual parsing.
  - **global or semantic context** defined by the topic of the document and consisting of semantic classes.
- Context features can be represented by relations with attributes.

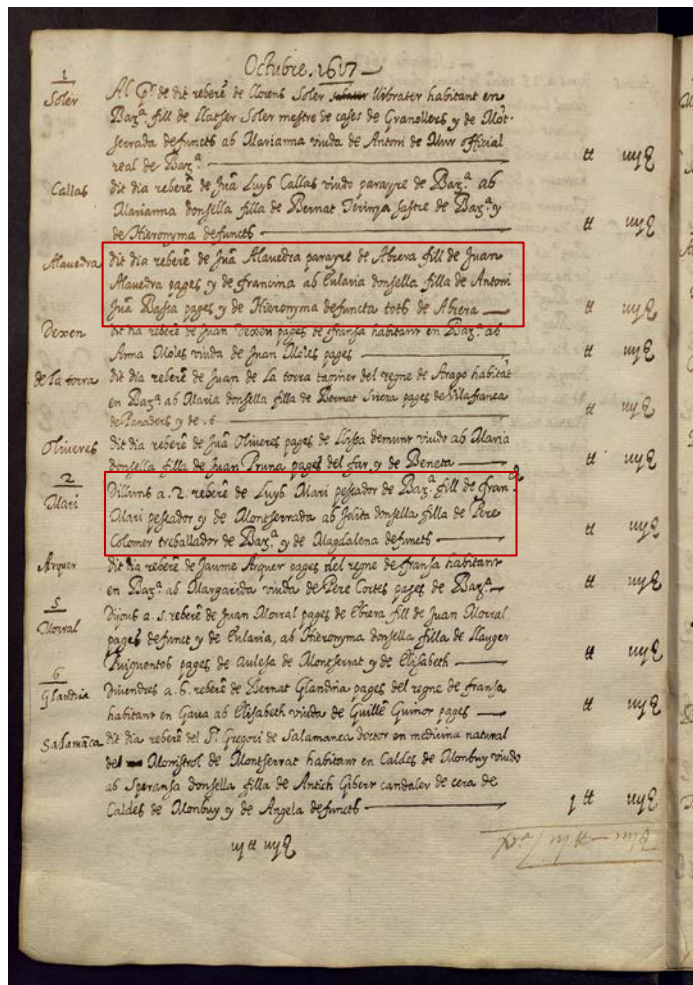
# Context: The structure of a marriage record

- Date (DI)
- Husband (HI)
- Husband's Parents (HPI)
- Wife (W)
- Wife's Parents (WPI)
- Key-words

Villans al p.<sup>r</sup> de Janer 1601 *reberem* de les esposat  
les de fraz Julia posador dela parroquia de Senare  
ras fill de Joa Julia pages y de ciuilia, *ab maria*  
donzella filla de ramon ferrer posador <sup>demaguer</sup> q.<sup>e</sup> y de  
Joana  
Die dia dehenam de Pera luea negociacione de i.<sup>a</sup> alama



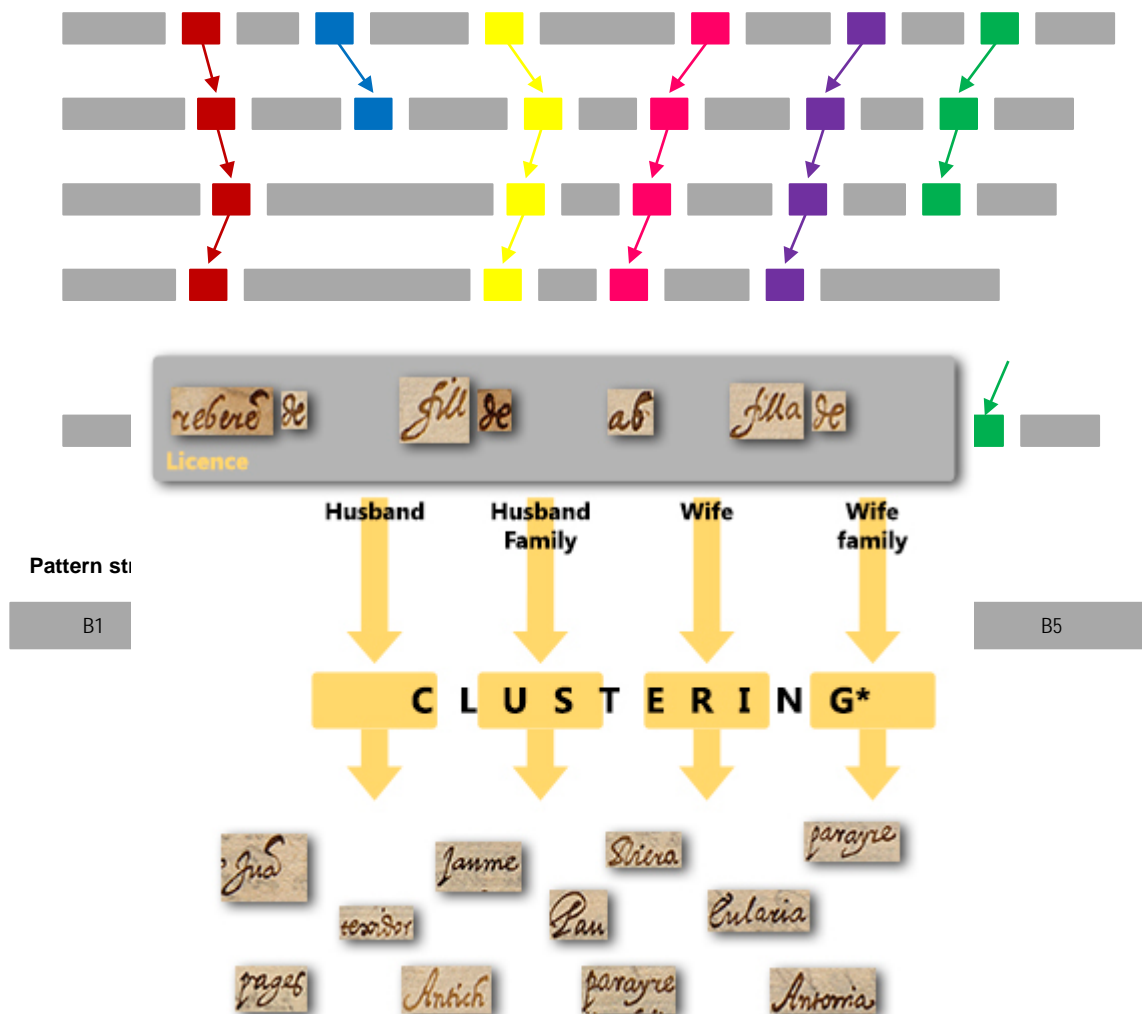
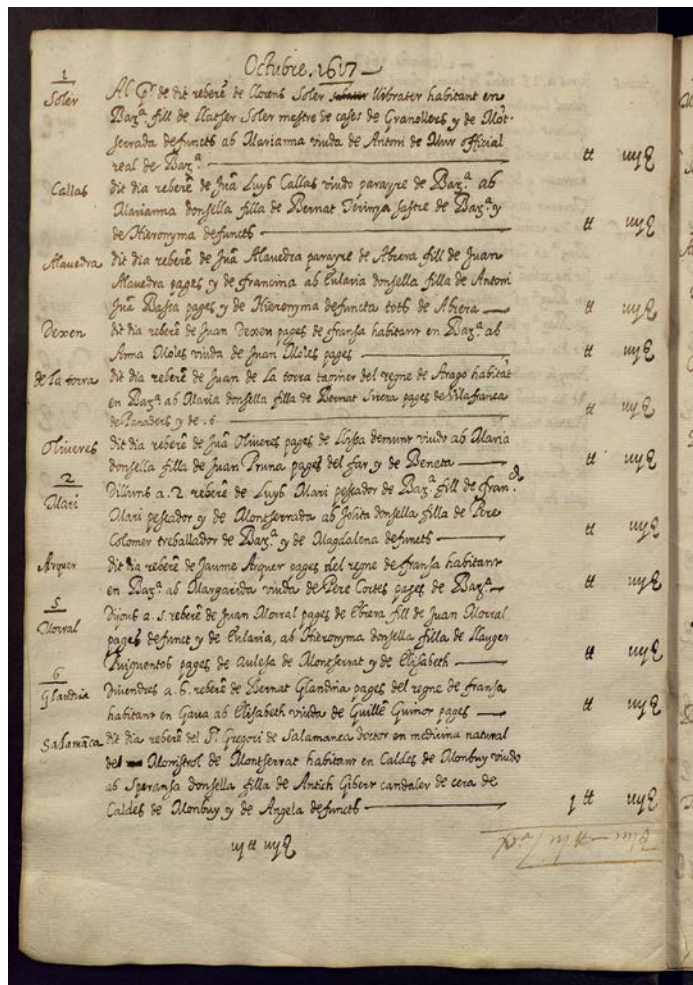
# Alignment of records



Dit dia rebere de Jua Alameda parayre de Abiera fill de Juan  
Alameda pages y de Francina ab Eulania donjella filla de Antoni  
Jua Bassa pages y de Hieronyma defuncta tott de Abiera

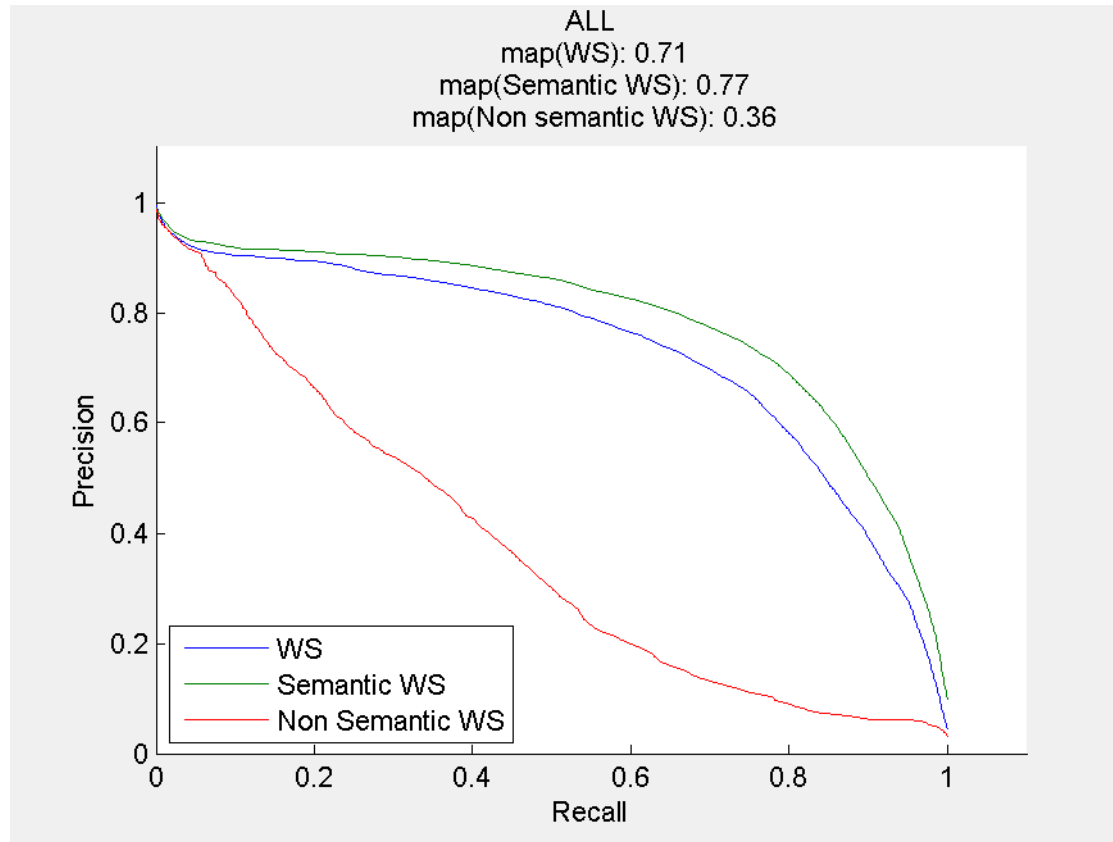
Dilluns a. 2. rebere de Luyb Alai peñador de Dag. fill de Fran.  
Alai peñador y de Montserrat ab Jolita donjella filla de Pere  
Colomer treballador de Dag. y de Magdalena defuncta

# Extraction of key and frequent words



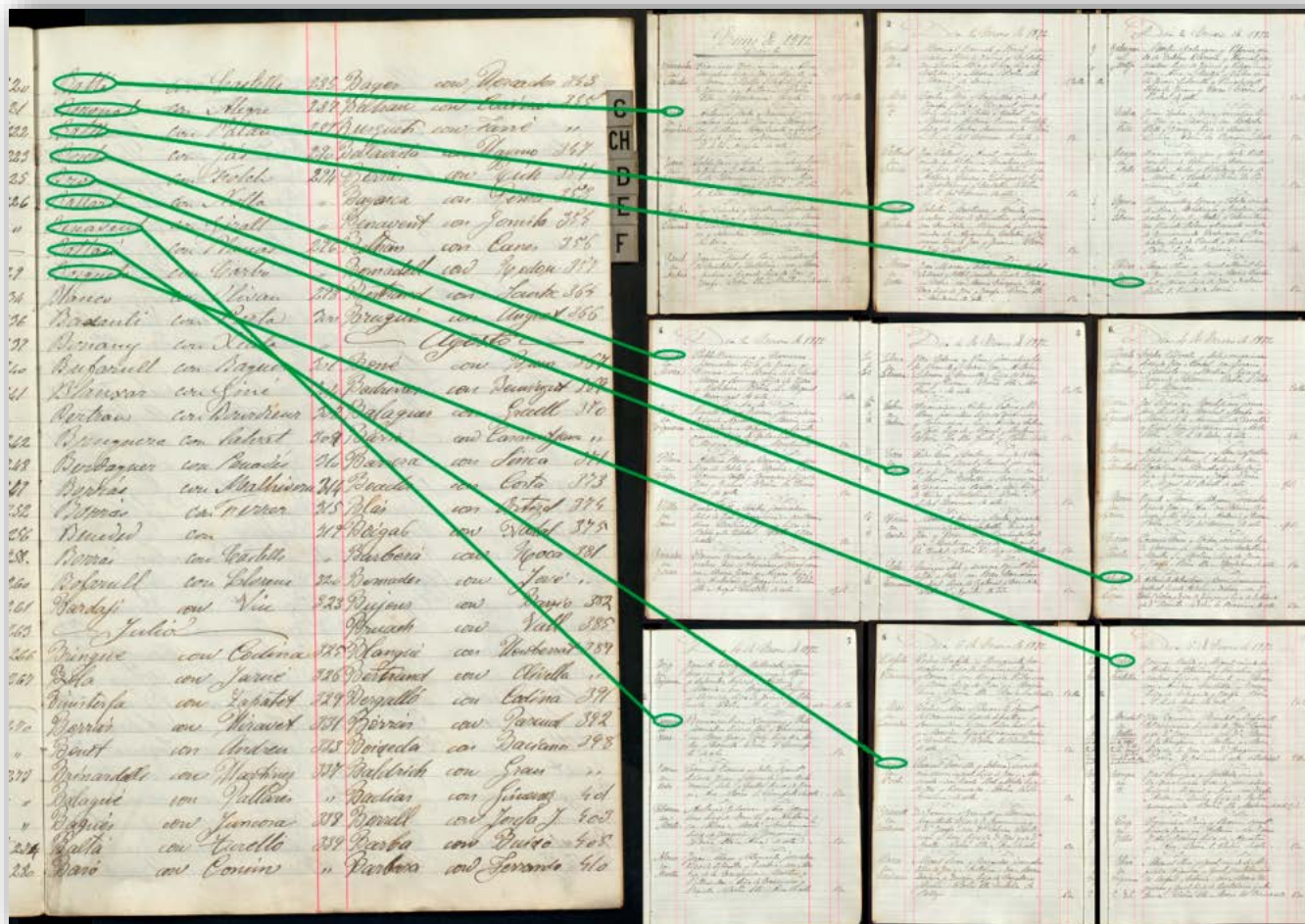
(\*) A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise In Second International Conference on Knowledge Discovery and Data Mining (1996), pp. 226-231 by Martin Ester, Hans P. Kriegel, Jörg Sander, Xiaowei Xu edited by Evangelos Simoudis, Jiawei Han, Usama Fayyad

# Results





# Alignment of handwritten words



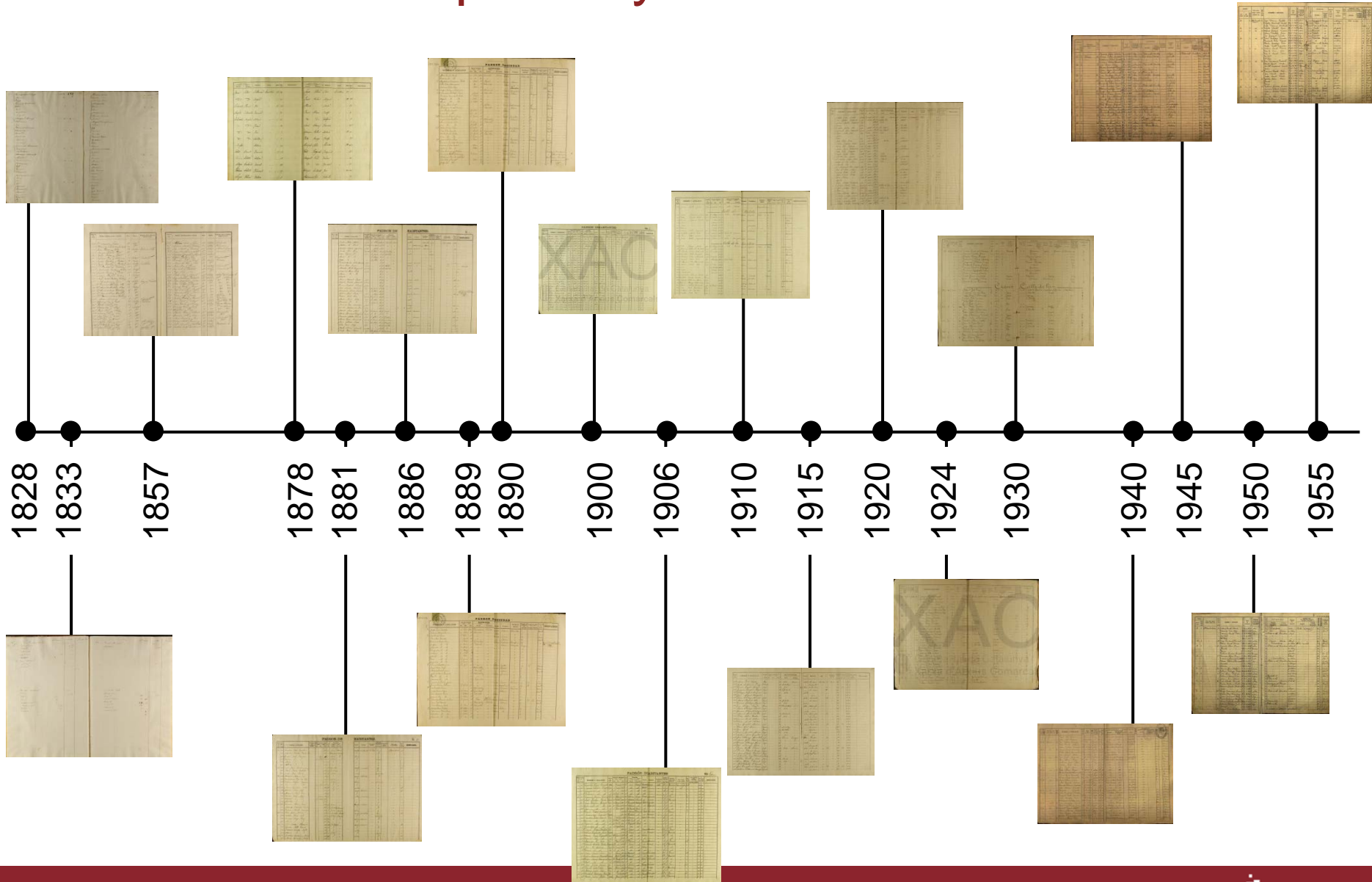
# Index

- Motivation and Context
- Word Spotting
  - Query by example
  - Query by string
- Context Aware Word spotting
- **Record linkage**
- Conclusions

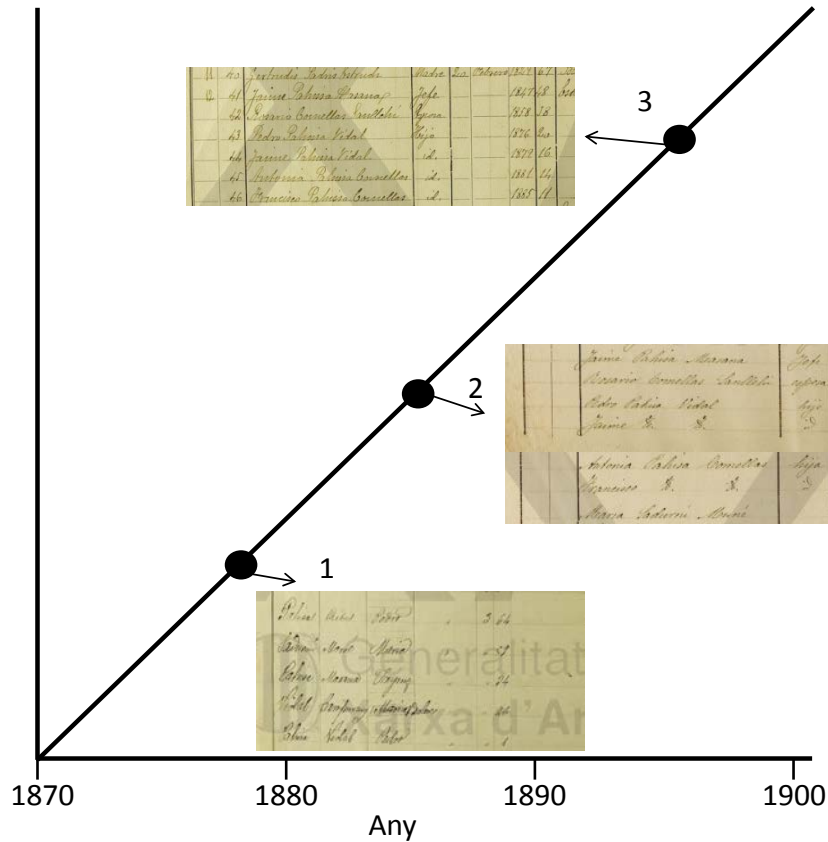




# Census Transcription by Information Transfer



# Individual life spans in census records

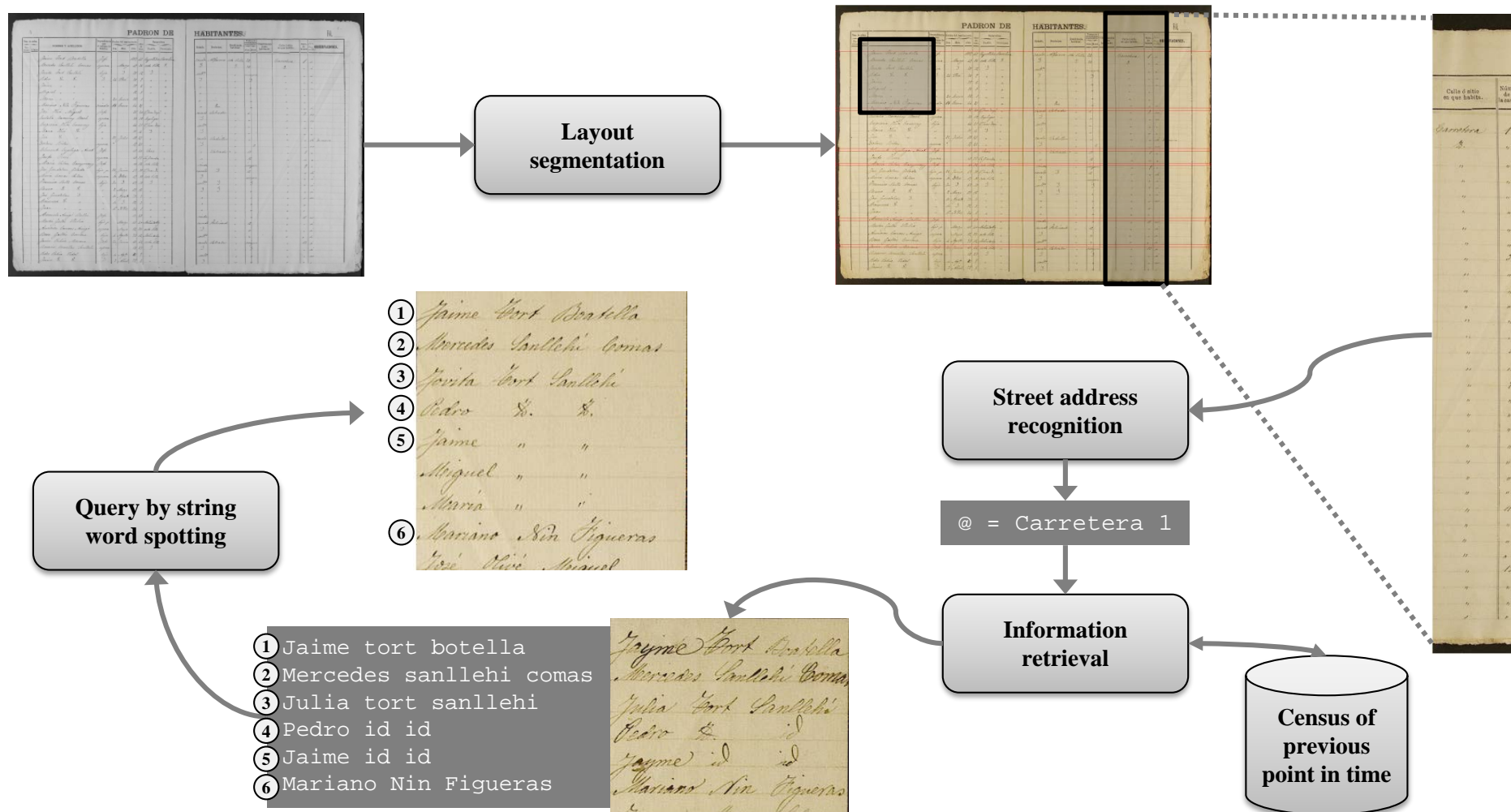


1 Padró 1878	Edat
Pedro Pahisa Ribas	64
Maria Sadurni Moner	59
<b>Jaime Pahisa Massana</b>	34
Dolores Vidal Campmany	26
Pedro Pahisa Vidal	1

2 Padró 1886	Edat
<b>Jaime Pahisa Massana</b>	42
Rosario Comellas Sanllehy	27
Pedro Pahisa Vidal	9
Jaime Pahisa Vidal	6
Antonia Pahisa Comellas	5
Francisco Pahisa Comellas	1
Maria Sadurni Moner	68

3 Padró 1896	Edat
<b>Jaime Pahisa Massana</b>	48
Rosario Comellas Sanllehy	38
Pedro Pahisa Vidal	20
Jaime Pahisa Vidal	16
Antonia Pahisa Comellas	14
Francisco Pahisa Comellas	11

# Extrinsic context in census records



# Index

- Motivation and Context
- Word Spotting
  - Query by example
  - Query by string
- Context Aware Word spotting
- Record linkage
- **Conclusions**



# Conclusions and future directions

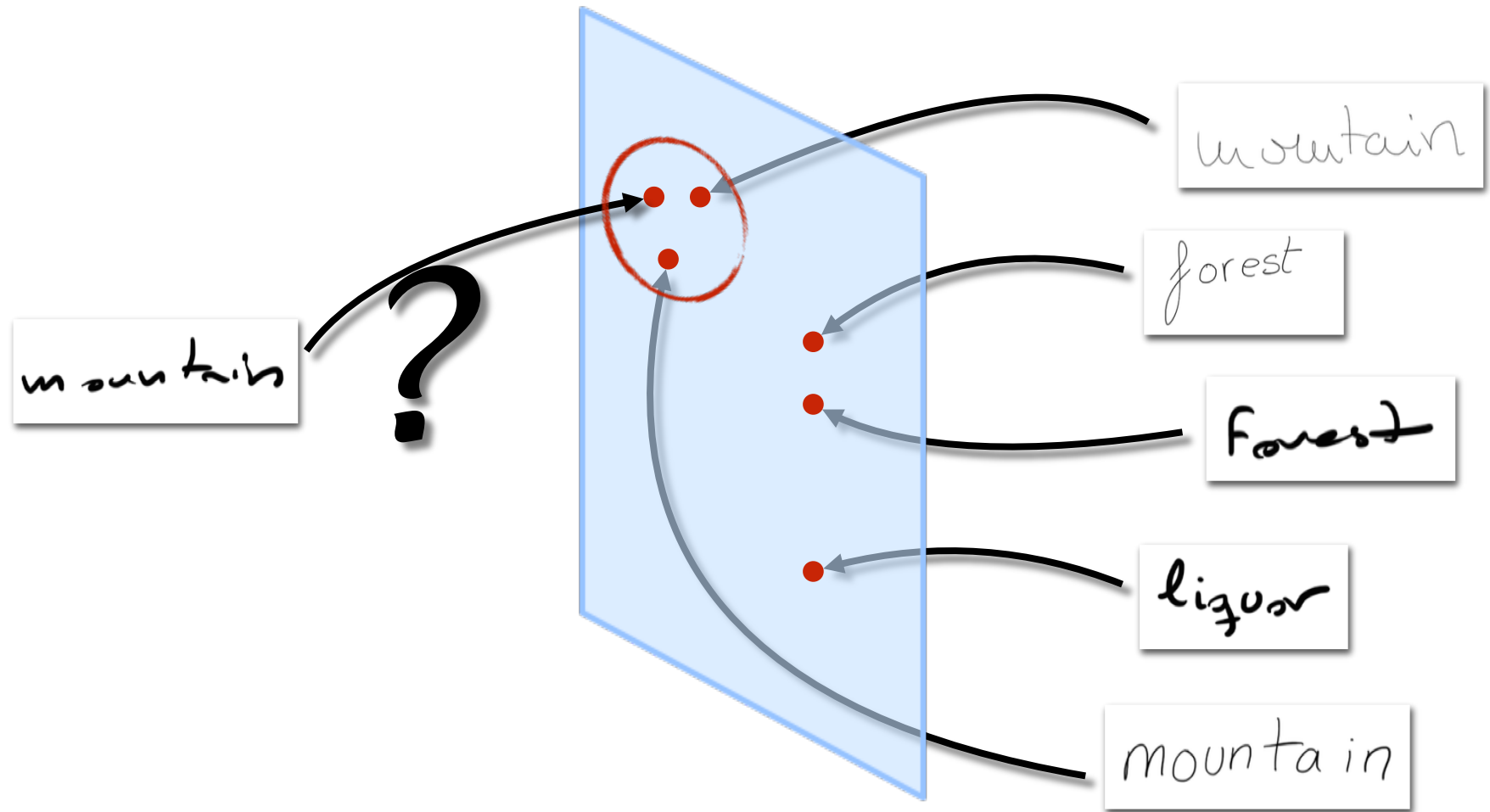
- Context aware word spotting
  - Grammars and language models
  - Spatial context
- In some fields (names, jobs, parishes, ...) there is a high frequency of a small number of classes (80-20%) → Take advantage of this so a month is transcribed and the spotting is “tuned” with the frequencies of words.
- Linkage between documents related to people (historical social networks).
- New paradigms for crowdsourcing: gamesourcing?



Thank you!

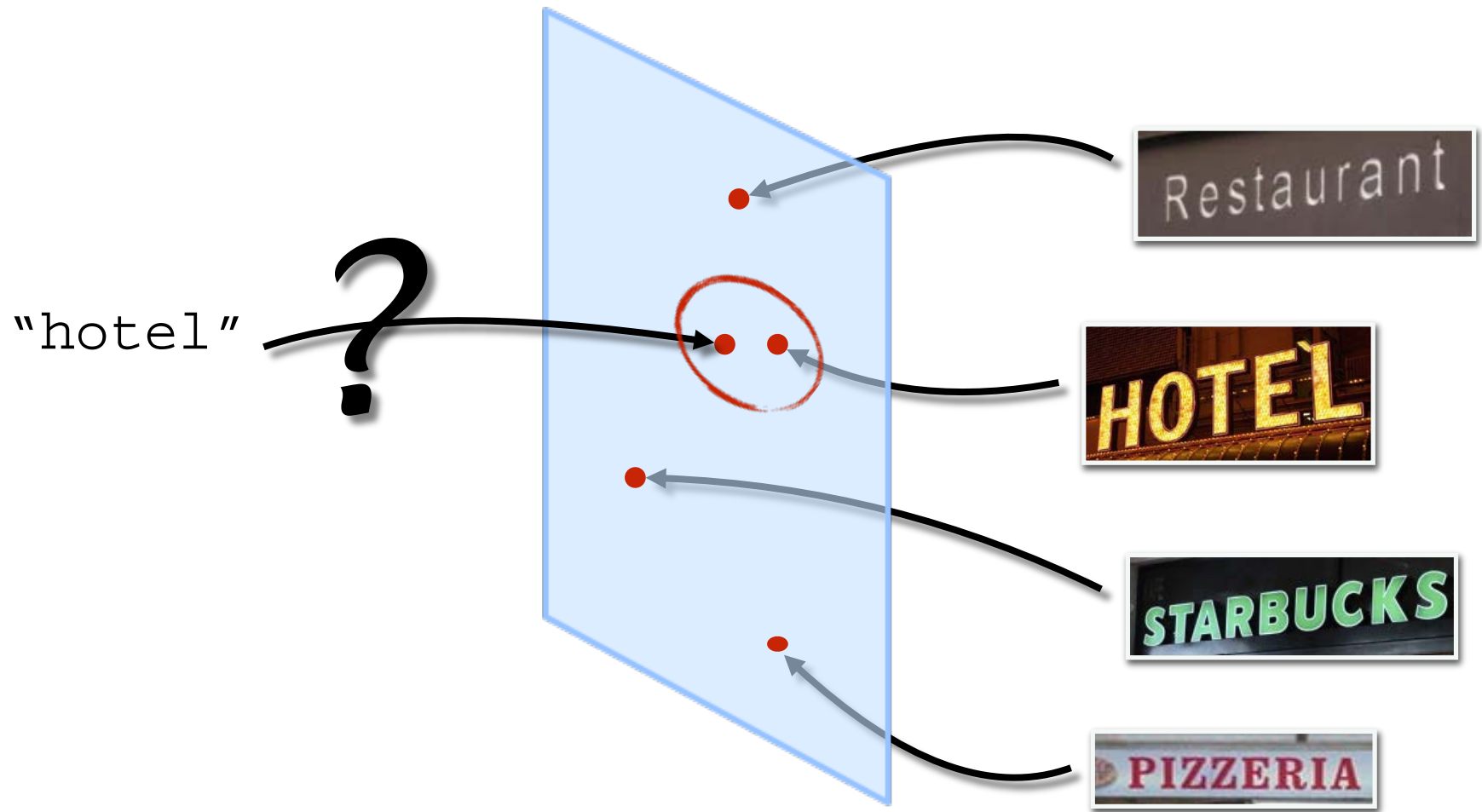


# Query by Example

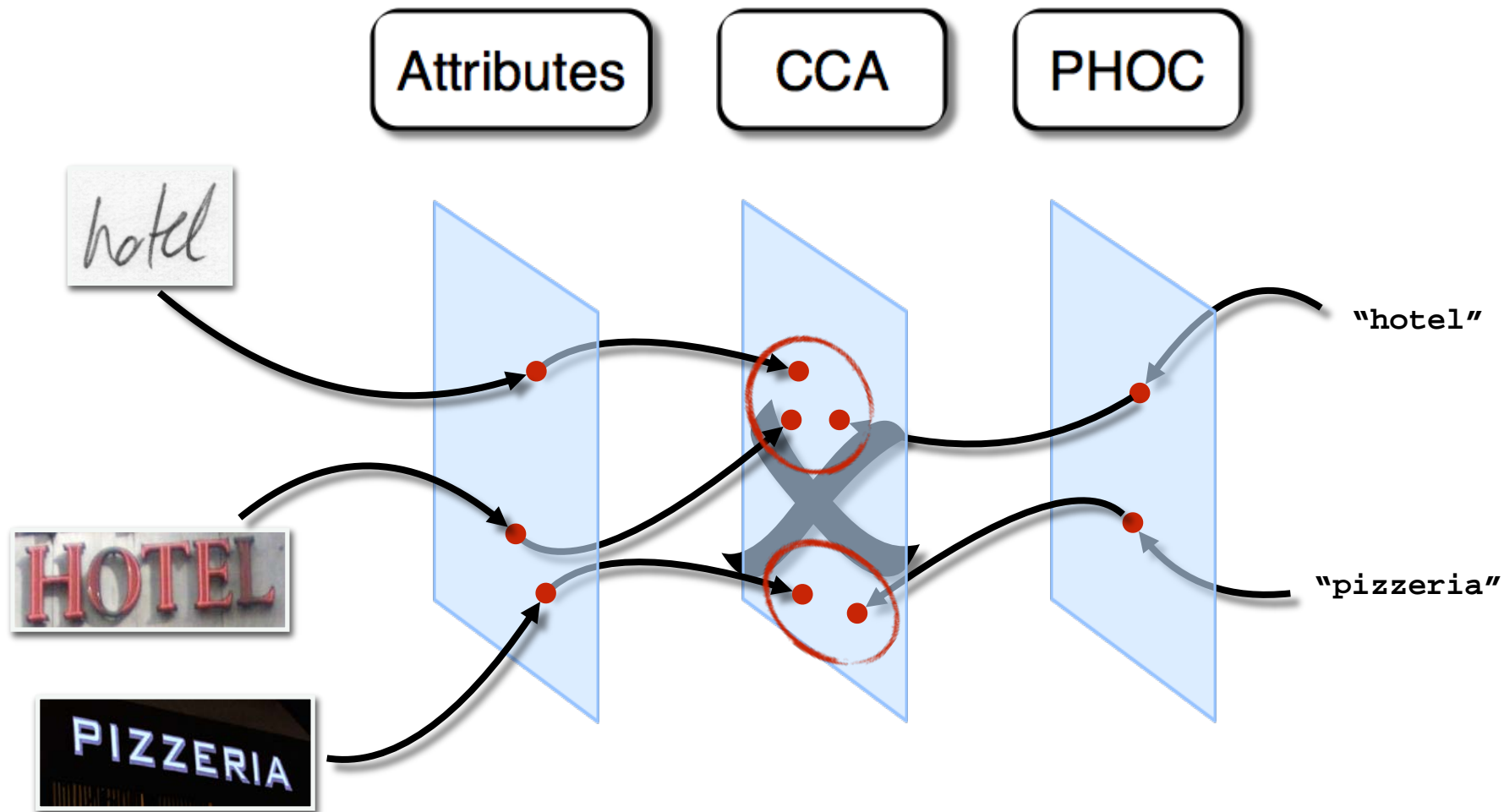




# Query by String



# Proposal



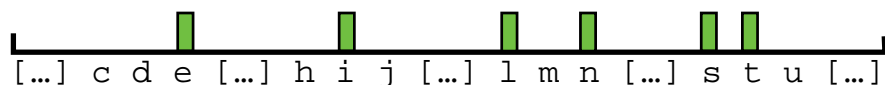


# Label Embedding

## Proposal: Pyramidal Histogram of Characters (PHOC)

Lvl 1

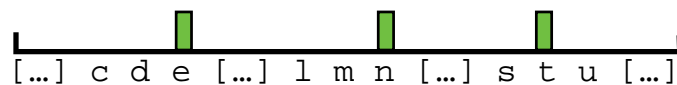
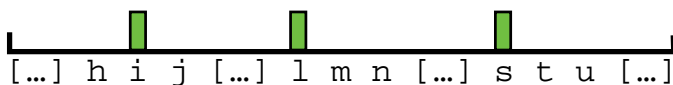
"listen"



Lvl 2

"lis"

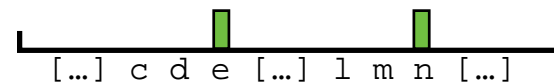
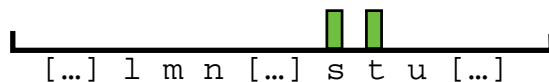
"ten"



Lvl 3 "li"

"st"

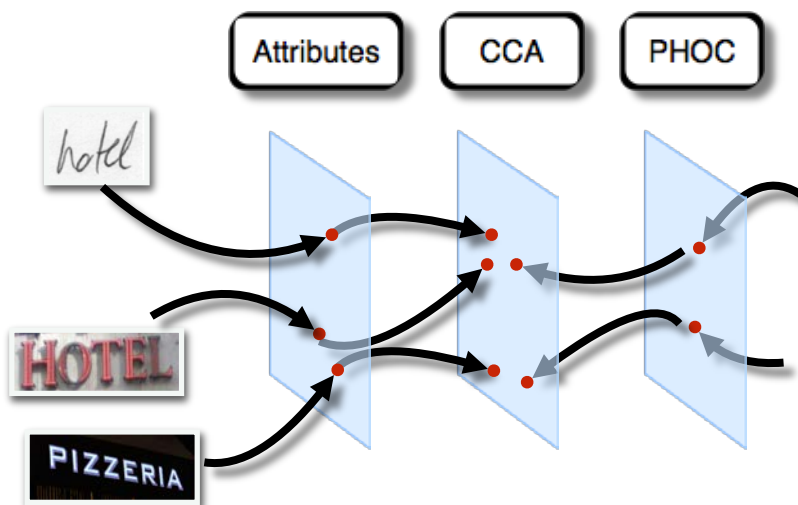
"en"



# Common Subspace

Canonical Correlation Analysis (CCA) to learn a common subspace between attribute scores and PHOCs

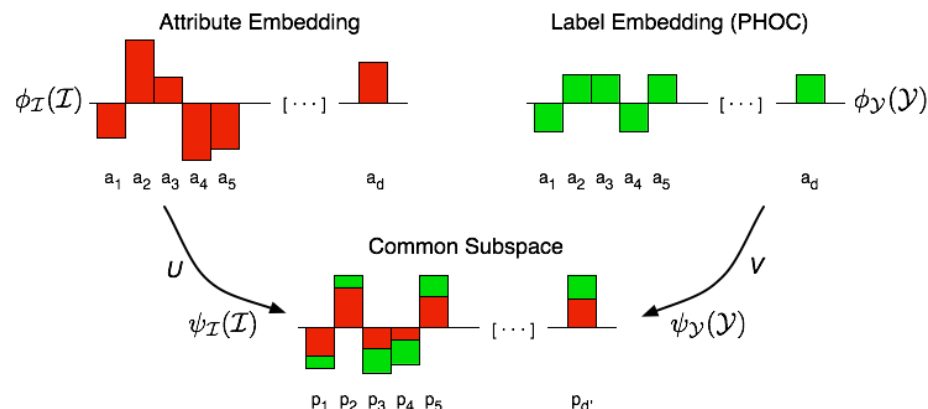
CCA finds a common space where the correlation of the projected views is maximal



# Common Subspace

Canonical Correlation Analysis (CCA) to learn a common subspace between attribute scores and PHOCs

CCA finds a common space where the correlation of the projected views is maximal



## Three advantages:

- Comparison between image and text embeddings is meaningful
- Attributes scores of images of the same word are brought together
- Dimensionality reduction (**96 dimensions**)